Essays in microeconomic theory

Schottmuller, C.

Publication date:
2012

Document Version
Publisher's PDF, also known as Version of record

Essays in Microeconomic Theory

Christoph Schottmüller

June 21, 2012

Essays in Microeconomic Theory

Proefschrift

ter verkrijging van de graad van doctor aan Tilburg University, op gezag van de rector magnificus, prof. dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in zaal DZ 1 van de Universiteit op donderdag 21 juni 2012 om 16.15 uur door

Christoph Schottmüller

geboren op 10 mei 1985 te Bad Bergzabern, Duitsland.

Promotiecommissie:

Promotor:	prof. dr. J. Boone
Overige Leden:	dr. C. Argenton
	prof. dr. E.E.C. van Damme
	prof. dr. P.J.-J. Herings
	prof. dr. B. Jullien
	prof. dr. H. Moreira

Acknowledgements

First of all, I want to thank Jan Boone. Jan, you made it possible for me to come to Tilburg as a side inflow candidate. You secured four years of funding for me by "stepping on a toe or two" (these are your own words from an email communication we had in early 2008 and it is–to the best of my knowledge–the only improperly referenced quote in this thesis). During those four years, you were an incredible supervisor. Not only did I learn a lot but we had quite some fun along the way. It was extremely helpful to coauthor papers with you and I very much appreciate that you were willing to do so. Especially (but not only) in the beginning, when we worked on what is now chapter 2 of this thesis, I would have been completely lost without your guidance. You taught me how to work on a research project. There was, however, something more important I could learn from you and I will simply call it a healthy attitude towards once own work and research in general. Also your support when it came to conferences, summer schools, visiting Toulouse and, of course, the jobmarket is highly appreciated. Last but not least, you also did an excellent job in doing what a supervisor is primarily meant to do which is giving comments and suggestions for my papers. Thanks for all your support and your enthusiasm.

The initial contact that brought me to Tilburg for my PhD was, however, an email I wrote to Eric van Damme. As soon as I arrived, Eric recruited me for Tilec and I think this had a profound positive impact on the thesis–especially the style it is written in. You also introduced me to the Dutch homecare market which eventually let to chapter 2 of this thesis. Since my work on chapter 2 initiated my work on the single-crossing property (see chapter 3 and 4), your influence on this thesis cannot be underestimated. Let me here thank you for your comments on the final thesis but also for your comments in earlier stages.

I want to thank Bruno Jullien not only for being in the thesis committee but also for hosting me in Toulouse. Chapter 4 was mainly written while I visited TSE. I benefitted

immensely from your comments and in some parts also from your earlier work on type dependent participation constraints. Thank you also for writing me a reference letter when I was on the jobmarket.

Chapter 4 would have been impossible without the earlier work by Humberto Moreira and coauthors. I am very grateful that you accepted to be on the committee and want to thank you for your extensive comments on all chapters of the thesis.

I am grateful to Jean-Jacques Herings for being on the thesis committee but also for playing an important role in my education as an economic theorist: In my first year, you were willing to teach an elective course in "applied theory" although I was the only student taking the course. I learned a lot in those sessions and want to thank you also for this.

I want to thank Cédric Argenton for being on the committee but also for his detailed comments on the thesis. I appreciated the discussions we had in the last years; those on my papers but also those on other issues. I am also grateful that you wrote me a reference letter when I was on the jobmarket.

I benefitted from comments by many other people. In no particular order, I want to thank Bert Willems (especially for suggesting the three type example in chapter 4), Matthias Lang, François Salanié, Florian Schütt and Jens Prüfer for comments and discussions. I also received helpful comments from seminar and conference participants at the CLEEN Workshop Amsterdam, NAKE Workshop Utrecht, Brazilian Workshop of the Game Theory Society São Paulo, Toulouse School of Economics, ENTER Jamboree Tilburg, Young Economist Spring Meeting Groningen, Summer School on Market Desing Louvain-la-Neuve, EARIE Stockholm, European Winter Meeting of the Econometric Society Tel Aviv, University of Southern Denmark Odense, Aalto University Helsinki, University of Copenhagen, Lund University, UT Sydney, Monash University Melbourne and Adelaide University. A special thank you to all my colleagues at Tilec and in the economics department of Tilburg University where I was allowed to present my work several times.

Last but not least, I want to thank all people who influenced this thesis in important but less direct ways: Administrative staff as well as my colleagues, friends, flatmates and family. I appreciate the support I got from you during my PhD time.

Contents

1 Introduction **1**

2 Procurement with specialized firms **7**

2.1 Introduction 7

2.2 Model 11

2.3 First best welfare monotone 20

2.4 First best welfare non-monotone 21

2.5 Robustness 27

2.5.1 Violation second order condition 27

2.5.2 Concavity in q 30

2.6 Conclusion 31

2.7 Appendix: Proofs 33

3 Health insurance without single crossing **45**

3.1 Introduction 45

3.2 Insurance model 49

3.2.1 Demand side model 49

3.2.2 Supply side 52

3.3 Income and health 56

3.4 Example 61

3.4.1 Duopoly 64

3.5 Conclusion 65

3.6 Appendix: Proofs 67

4 Adverse selection without single crossing **71**

4.1 Introduction 71

4.2 Examples 73

4.2.1 Example settings where single crossing is violated 73
4.2.2 Three type example 77
4.3 Literature 79
4.4 Model 82
4.5 Optimal contract 88
4.5.1 Necessary conditions 88
4.5.2 Monotone solution 90
4.5.3 Continuous solutions 98
4.5.4 Distortion at the top 102
4.5.5 Stochastic contracts 104
4.6 Discussion 105
4.7 Conclusion 107
4.8 Appendix 109
4.8.1 Variational condition 109
4.8.2 Proofs 110
4.8.3 Existence of an optimal contract 124

5 Cost incentives for doctors **127**
5.1 Introduction 127
5.2 Formal setting 130
5.3 A simple example 131
5.3.1 No cost incentives 132
5.3.2 Cost sensitive doctor 133
5.3.3 Variation I: Restricting the choice set 134
5.3.4 Variation II: Increasing costs 134
5.4 Model and results 135
5.5 Discussion and conclusion 142
5.6 Appendix: Proofs 144

Chapter 1

Introduction

Economics is often described as the social science dealing with the allocation of goods and services and the efficiency of this allocation. The market mechanism has been the most prominent allocation mechanism in this framework. In terms of modern economics, this prominence is rooted in general equilibrium theory. The main insight is the first welfare theorem which states that the market mechanism leads to efficient allocations under certain assumptions.

One crucial assumption of the welfare theorem is price taking behavior, i.e. no player is big enough to influence the market price with his behavior. This means, for example, that a firm cannot strategically restrict its output to increase the market price. However, as Coase (1960) points out, efficient allocations should even be expected in bargaining markets with few players. The idea is that it would always be efficient to switch from an inefficient to an efficient allocation while sharing the additional rents from this switch.

The Coase argument, however, hinges on the assumption of symmetric information, i.e. every player knows not only his own valuation of any possible allocation but also the valuation of all other involved parties.[1] Akerlof (1970) shows that even markets with a large number of buyers and sellers are no longer efficient if this assumption fails. He gives the example of the used car market and argues that sellers of used cars are better informed about the quality of the car than buyers. As a buyer cannot distinguish good from bad cars, he is only willing to pay an average price. It might very well be that the owners of high quality cars find this average price too low to sell their car. Consequently, high quality cars are not sold even if a sale was efficient, i.e. buyers value a high quality used car more than sellers. This is an illustration of the adverse selection principle: At any given market price the sellers of the low quality cars will be more eager to sell their

[1] The argument of Coase depends, of course, on other assumptions as well, e.g. clearly defined property rights and absence of transaction costs. However, these topics will not be dealt with in this thesis.

cars than the sellers of the high quality cars.

The idea that asymmetric information leads to inefficient allocations was also shown in settings with few players. Mirrlees (1971) analyzes a model where a government sets an income tax schedule to maximize its redistributive objectives. The assumption here is that a government can only observe the income someone has but not the effort this person had to exert to achieve this income. If people differ in their productivity, it is impossible to achieve full equality: A productive worker could always claim to be unproductive. Because of his high productivity he has to exert less effort than an unproductive worker to reach the same income. Mirrlees (1971) shows how the government optimally distorts the labor supply decision of the workers. The main results are that marginal tax rates vary between 0 and 1 where only the most productive worker faces a marginal tax rate of 0. This implies that all but the most productive worker work less than socially efficient as taxation distorts their labor supply decision.

The work of Mirrlees is especially important because it presents the technical machinery that has been used in many applications of similar models later on. Examples are models of non-linear pricing (Mussa and Rosen, 1978), regulation (Baron and Myerson, 1982) or insurance (Stiglitz, 1977). All these models look at a principal (monopolist, regulator, monopolist insurance) who chooses a mechanism according to his preferences. He offers a menu from which the agent (consumer, regulated firm, insured) can choose his preferred action (quality, quantity, insurance coverage). By his choice, the agent reveals his private information (willingness to pay, efficiency level, risk). The common assumptions are that the principal can commit to the offered menu, i.e. he cannot retract his offer and substitute it with a different one after the agent has chosen from the menu. By allowing the principal to make a take-it-or-leave-it offer, the principal has full bargaining power and only the presence of asymmetric information allows the agent to earn rents. All these models share the properties of the Mirrlees (1971) model: The allocation for all but the best type is distorted below first best.

The main technical machinery used in these models is the following. The private information of an agent is denoted by a one dimensional parameter called "type". The first step is using the revelation principle. This principle says that any equilibrium of any mechanism can be implemented through a direct revelation mechanism. It is therefore sufficient to concentrate attention on direct revelation mechanisms. A direct revelation

mechanism consists of an allocation rule which assigns an allocation to each possible type. Then the agent is asked for his type and the allocation corresponding to his announced type is implemented. One has to make sure that each type has an incentive to truthfully reveal his type. This last requirement is known as incentive compatibility constraint or incentive constraint for short.

Technically, it would be very difficult to consider the incentive constraints between all types. However, Mirrlees (1971) introduced assumptions that greatly simplify this task and have–for this reason–been part of the literature ever since. It is assumed that higher types are "better" in the following sense: A higher type has (i) a higher utility from any possible contract and (ii) a higher marginal rate of substitution between the decision and money at any possible contract ("single crossing"). Let me give an example to illustrate this: In a regulation setting, a higher type firm would have lower costs and lower marginal costs at any output level. Hence, type denotes an efficiency level and the private information of the agent basically conveys information concerning "how good" he is. This simplifies the problem as it can be shown that non-local incentive constraints can be neglected under the assumptions above. Put differently, if each type prefers his contract to the contracts of very close by types, he will also prefer his contract to the contracts of far away types.

Three chapters of this thesis relax the assumption above. In chapter 2, a procurement auction is analyzed in which firms are specialized. Specialization directly violates part(i) of the assumption above: We assume that low types are relatively efficient for low quality production but high types are relatively efficient for high quality production. Hence, a firm's private information is no longer its efficiency but its production technology. At the quality level a firm is specialized in, it produces at lower costs than any other firm. While such setups have already been analyzed in settings with one principal and one agent, the extension to an auction setting is new and creates technical challenges as well as new insights. For example, non-local incentive constraints have to be checked and normal scoring rule auctions can no longer implement the optimal mechanism.

Chapter 3 analyzes a health insurance setting where the risk of falling ill is private information of the agent. In such a setting, part (ii) of the assumption above would mean that someone with a higher probability of falling ill is willing to pay more for an increase in insurance coverage. While this seems reasonable on first sight, it is not obvious in the

health context: It is well documented that high health risk is strongly correlated with low income. It is also documented that, at partial coverage, people with low income often forego treatment when falling ill in order to save copayments. But then it is not clear that someone with a higher risk but lower income will be more eager to buy insurance coverage.[2] At full coverage, however, income does not play a role for utilization and the standard intuition that higher risk agents are more eager to buy insurance coverage prevails. It is shown that this setup–with a violation of part (ii) of the assumption above–can help to explain the empirical puzzle that people with low health risks tend to have high insurance coverage.

Chapter 4 deals with a violation of part (ii) of the assumption above in a general principal agent setting with quasilinear utility functions. The focus is here how to deal with binding non-local incentive constraints. It is shown that binding non-local incentive constraints reduce the usual distortion in those models and can even lead to upward instead of downward distortion. The latter distortion can even happen at the top type. A rough intuition is that the normal downward distortion enables the principal to extract more rents from high types. If this leads to a violation of a non-local incentive constraint, the principal has extracted too much from high types: High types wants to claim that they are (very) low types. In short, the principal cannot extract that much from high types and therefore the motive for the downward distortion is counteracted.

The last chapter of this thesis turns to a different kind of setting in the theory of asymmetric information. Following Crawford and Sobel (1982), a cheap talk framework is used to analyze communication between doctor and patient. The question asked is whether welfare is maximized if a doctor takes the costs of the health insurance into account when deciding about a prescription. The alternative considered is that the doctor maximizes the patient's utility without taking costs to the insurance into account. At first glance, internalizing costs seems to increase welfare. However, there is a problem. If the doctor takes these costs into account his objective differs from the patient's objective, i.e. the patient would favor more expensive treatment because his insurance pays for it. To get the more expensive treatment, the patient will try to convince the doctor that he

[2]Note that from the insurance point of view it does not matter whether someone is unable or unwilling to buy more coverage. However, we assume in chapter 3 that the provision of full insurance is first best efficient.

is an especially suffering case by exaggerate his symptoms. The doctor will anticipate this behavior but communication in equilibrium will be noisy. Worse communication makes it more difficult for the doctor to get the right diagnosis and therefore welfare is decreased by this communication effect. Whether the communication effect or the cost internalization effect dominates depends on the specific setting.

Chapter 2

Procurement with specialized firms[3]

2.1. Introduction

The literature on procurement and optimal incentive regulation, see for example Laffont and Tirole (1987), Laffont and Tirole (1993) or Che (1993), assumes that firms have private information with regard to its cost function. As usual in screening models, this private information is represented by a "type" which is assumed to be a scalar. It is then assumed that higher types are better in the sense that higher types have lower costs. If costs, for example, depend on the quality produced, this means that a higher type has lower costs for any quality level.

On the other hand, the private information of a firm is often interpreted as the production technology it uses. This technology was determined in the past and can therefore be treated as given in the context of one specific procurement contract. Following this interpretation, one should expect that firms chose production technologies that are not obviously inferior to alternative technologies, i.e. there should be some level of quality for which the technology of a firm is efficient. Put differently, firms are specialized in the production of a certain quality level. As we argue below, this specialization is not covered by standard procurement models which assume that higher types are better in order to obtain a single crossing property.

Put differently, the procurement literature so far focusses on private information concerning absolute efficiency of a firm. We will think of private information in terms of production technology and specialization.

To illustrate what we mean by that, we describe the market for home care in the Netherlands which was recently liberalized. Local governments now procure home care

[3]This chapter is based on Boone and Schottmüller (2011b).

for their citizens and money saved on the procurement can be used now by local governments for other things, like sports facilities (that is, the money received from the central government to pay for home care is not earmarked). However, the local government does have a duty to provide care of some minimum standard. It used to be the case that regional care offices procured without much incentive to save money. Due to this liberalization, new players have entered the market. For example, cleaning companies considered moving into home care. As these new players have no experience with care (to illustrate, they did not use to hire nurses or other professionals with a medical background), they are seen as low quality players. At this low quality level, however, they are cheaper than traditional firms. That is, they can provide simple services like house cleaning and shopping more cheaply than incumbent home care companies. In this sense, incumbents are specialized in high quality production while entrants are specialized in low quality/low costs production.[4]

This pattern –where incumbents are specialized in high quality service while entrants specialize in a low quality (low price) service– is typical after liberalization. Many European countries have liberalized sectors like post, taxis, air transport, railway or local transport. This has led to entry by players who offer lower quality in, for instance, the following sense: only make deliveries twice a week (instead of 6 days a week), drive cars substantially cheaper than a Mercedes (see `http://www.tuktukcompany.nl/` for an example), operate planes with reduced seat pitch and limited on board service as well as offering less connections, use old trains and buses to transport people. A reaction often heard by customers and/or incumbents is that the liberalization is bad for welfare because of the lower quality.[5]

[4]To a certain extent this can be resolved through market separation in *care* and *support*. People who do not need medical attention but only someone to clean their home, can be served by cleaning companies. While patients who stay at home and need a nurse can be served by the incumbents. Hence at the extremes of the home care spectrum, market separation can alleviate the issue. However, many cases in home care are not so clear cut. To illustrate, a nurse helping an elderly woman putting on her clothes in the morning and cleaning the house may recognize the first signs of dementia that would be overlooked by an employee of a cleaning company. In such a separated market, what we write above applies to the *support* segment of the market.

[5]One could even take a broader point of view and also consider the case of foreign workers entering a domestic labor market. In the EU there was a heated debate about Polish workers coming to the west in case Poland would join the EU. Again some people argued that this is bad since Polish workers are

In each of these cases, one could argue either that quality did not decrease at all or that before liberalization quality was inefficiently high. In the former case, incumbents spread rumors to reduce the probability that entrants win contracts. In the latter case, after liberalization quality goes down but total surplus rises. Presumably, in some of the examples mentioned either of these two cases arise. However, we are interested in the case where indeed entrants offer both lower quality and lower total surplus than incumbents.[6] The question we ask is: How should a planner (in the home care example: the municipality) who wants to maximize welfare optimally organize the procurement in the face of such low quality entrants?

We show the following results. Think of low (high) type firms as firms specialized in low (high) quality production. First, if low types (e.g. entrants in the examples above) are indeed worse high types (incumbents in the examples) with respect to first best welfare, the incumbents (under the optimal procurement rules) do not lose from entry. Second, only if first best welfare first decreases and then increases in type, types specialized in high quality can lose in the following way: A low quality provider (entrant) can win the procurement even though the high quality provider (incumbent) would provide higher welfare under the optimal procurement rules. Third, in this latter case, quality is distorted above first best for some types and below first best for others. Fourth, in both cases an interval of types has zero profits ("profit bunching"). Although all types in this interval have zero profits, they produce different qualities when winning the contract. Put differently, a mass of types will have no economic rents under the optimal contract although types are perfectly separated in equilibrium. Fifth, if first best welfare is monotone in type, relatively simple auctions can implement the optimal mechanism.

Technically speaking, a contribution of the paper is to solve a two-dimensional mechanism design problem. A technical challenge is that local incentive compatibility is not straightforwardly sufficient for non-local incentive compatibility, i.e. non-local incentive constraints have to be checked explicitly. To illustrate the problem, view profits as a function of the probability of getting the contract. The assumption that firms are spe-

supposedly less qualified than domestic workers.

[6]There are two reasons for this focus. First, as argued below, in the home care example mentioned above there is evidence suggesting that the entrants offer lower surplus. Second, if the entrants offer higher surplus than the incumbents, it is clear that they should be used by a value maximizing planner. Further, if the cheaper entrants offer higher surplus, no service should be procured from the incumbents.

cialized implies then that "marginal profits" (where marginal refers to a slightly higher probability of getting the contract) are not monotone in type. This is equivalent to a violation of single crossing in one dimensional models. As is well known, non-local incentive compatibility does not follow from local incentive compatibility if single crossing is not satisfied.

Our paper is related to the literature on procurement, especially to those papers in which more than price matters, e.g. Laffont and Tirole (1987), Che (1993), Branco (1997), Asker and Cantillon (2008) or Asker and Cantillon (2010). This literature shows how quality (or quantity) is distorted away from first best for rent extraction purposes. It also analyzes how simple auctions can implement the optimal mechanism. These papers assume that firms are not specialized, i.e. higher types have lower costs for all quality levels. This assumption seems to be too strong in many settings, e.g. newly liberalized industries. We show that relaxing it leads to zero economic rents for a mass of types which is, to our knowledge, a new result in the literature on procurement auctions. We also show that implementation of the optimal mechanism by standard auctions, e.g. scoring rule actions, is no longer straightforward when firms are specialized.

Our paper connects the literature on competitive procurement with the literature on countervailing incentives, see Lewis and Sappington (1989) for the seminal contribution and Jullien (2000) for the most general treatment. By assuming that firms are specialized, our paper uses a cost function that resembles the utility functions of the countervailing incentives literature. Our result that the participation constraint is binding for a mass of types is also typical for this literature. We contribute by allowing for several agents bidding for the contract while the countervailing incentive literature focuses on settings with one principal and one agent. As a consequence of this one agent setting, the probability of being contracted is one for the agent. Consequently, local incentive compatibility constraints are sufficient for non-local incentive compatibility and many of the technical challenges encountered in our paper do not occur. From an applied point of view, having more than one firm leads to the result that optimal procurement auctions can be second best inefficient.

An exception from the focus on one agent in the countervailing incentives literature are papers on auctions with valuation dependent externalities, see Carrillo (1998), Figueroa and Skreta (2009) or Brocas (2011). The outside option is type dependent

since the agent suffers the externality even if he does not participate. The main difference between our paper and this literature is the existence of another variable, i.e. quality in our paper, while the auction literature focuses on the problem of allocating one exogenously given good. Hence, the only variable is the probability of getting the good. Since the used preferences satisfy single crossing, non-local incentive constraints play again no role.

As we solve a mechanism design problem with two variables, i.e. quality and the probability of winning, our paper is also related to the literature on multidimensional screening as surveyed in Rochet and Stole (2003). We contribute here by analyzing a two-dimensional screening model with countervailing incentives. Other screening models with one-dimensional type and multidimensional decisions include, for example, Matthews and Moore (1987) or Guesnerie and Laffont (1984). In these papers, single crossing is assumed in each dimension which rules out the specialization we have in mind.

The set up of the paper is as follows. In section 2.2, we present the model. Section 2.3 analyzes the case where first best welfare is monotonically increasing in type while section 2.4 deals with U-shaped first best welfare. In the latter case we find a discrimination result, i.e. some types with lower second best welfare are preferred to types with higher second best welfare. Section 2.5 shows how the model extends to situations in which the assumptions of section 2.2 are not met and section 2.6 concludes. Proofs are relegated to the appendix.

2.2. Model

We consider the case where a social planner procures a service of quality $q \in Q \subset \mathbb{R}_+$ where Q is a convex set. The gross value of this service is denoted by $S(q)$ where we normalize quality in such a way that $S(q) = Sq$ for some $S > 0$.[7] The cost of production is denoted by the three times continuously differentiable cost function $c(q, \theta)$ where a firm's type θ is private information of the firm. We assume that each firm's type is drawn independently from a distribution F on $[\underline{\theta}, \bar{\theta}]$ which has a strictly positive density f. We also assume that c is (at least) three times continuously differentiable.

[7] This is, given our assumptions on the cost function, without loss of generality for weakly concave gross values $S(q)$.

We make the following assumptions on the cost function c and distribution function F.

Assumption 2.1. *We assume that*

- *the function* $c(q,\theta)$ *satisfies* $c_q, c_{qq} > 0, c_{q\theta} < 0, c_{\theta\theta} \geq 0$,
- *for* $q \in Q$ *it is the case that* S *is high enough compared to* $c(q,\theta)$ *that the planner always wishes to procure (regardless of the type realization) and*
- *the function* F *satisfies* $\frac{d((1-F(\theta))/f(\theta))}{d\theta} < 0$ *and* $\frac{d(F(\theta)/f(\theta))}{d\theta} > 0$.

These assumptions are standard in the literature. The first part says that c is increasing and convex in q. Higher θ implies lower marginal costs c_q (the Spence-Mirrlees condition) and c is convex in θ. It will become clear that this convexity is part of the idea of specialized firms. The second assumption formalizes the idea in our home care application that the government cannot decide not to provide the service. That is, it is always socially desirable for the service to be supplied. The third part is the monotone hazard rate (MHR) assumption. Usually this assumption is only made "in one direction". However, in the literature on countervailing incentives it is standard to have MHR "in both directions", see for example Lewis and Sappington (1989) or Jullien (2000). Well known distributions that satisfy MHR include normal, uniform and exponential distributions.[8] In section 2.5, we discuss what happens if MHR is not satisfied.

The following assumption states that firms are specialized which is the case we want to analyze in this paper.

Assumption 2.2. *For each* $\theta \in [\underline{\theta}, \bar{\theta}]$, *there exists* $k(\theta) > 0$ *such that*

$$c_\theta(q,\theta) \begin{cases} > 0 & \text{if } q < k(\theta) \\ < 0 & \text{if } q > k(\theta) \end{cases}$$

Further,

$$c_{q\theta\theta}(q,\theta) \begin{cases} \leq 0 & \text{if } q < k(\theta) \\ \geq 0 & \text{if } q > k(\theta) \end{cases}$$

$$c_{qq\theta}(q,\theta) \begin{cases} \geq 0 & \text{if } q < k(\theta) \\ \leq 0 & \text{if } q > k(\theta) \end{cases}$$

[8] See Bagnoli and Bergstrom (2005) for a more complete overview.

Hence, for high values of q, a higher type θ produces q more cheaply. This is the usual assumption. We allow for the possibility where low values of q are actually more cheaply produced by lower θ types. To illustrate, high θ incumbents may have invested in (human) capital that makes it actually relatively expensive to produce low quality. If the quality of the product is mainly determined by the qualification of the staff, incumbents might have more expensive but also more qualified workers. Replacing these workers is, especially in Europe, costly because of labor market rigidities and search costs. Consequently, it is more expensive for incumbents to produce low q than for entrants (and the other way around for high q). The function $k(\theta)$ is implicitly defined by $c_\theta(k, \theta) = 0$. By assumption 2.1, $k(\theta)$ is differentiable and monotonically increasing. Put differently, as θ increases the quality level $k(\theta)$ where $c_\theta = 0$ (weakly) increases.

In some sense, our assumption that c_θ switches sign in q follows naturally from the sorting condition $c_{q\theta} < 0$. However, it is the main departure from the existing literature on procurement which assumes $c_\theta < 0$ or equivalently that $k(\theta) < 0$ which implies that $c_\theta < 0$ in the relevant domain. Put differently, the existing literature assumes that types can be ranked in terms of efficiency irrespective of q. We allow efficiency advantages to depend on q and therefore firms can be specialized in producing a certain quality.[9]

To make sure that (i) the planner's objective function is concave in q and (ii) quality q increases in type, it is standard in the literature to make assumptions on third derivatives $c_{q\theta\theta}, c_{qq\theta}$. If c_θ does not switch sign, the usual assumption is that these derivatives should not switch sign either. This is different in our case. To ease the exposition we make the assumptions on the third derivatives above and discuss in section 2.5 what changes if these assumptions are not satisfied. Note that we allow for the simple case where these third derivatives are equal to zero.

As c_θ can be positive, it is not clear how first best welfare varies with θ. Below we define the two cases that we consider here. In order to do this, we introduce the following notation. First best output is defined as

$$q^{fb}(\theta) = \arg\max_q Sq - c(q, \theta) \tag{2.1}$$

which is a singleton as $c_{qq} > 0$ by assumption 2.1. First best welfare is denoted by

$$W^{fb}(\theta) = Sq^{fb}(\theta) - c(q^{fb}(\theta), \theta). \tag{2.2}$$

[9] If q is interpreted as quantity, we allow firms to be specialized in a certain scale of production.

Our final assumption makes sure that we can focus on two relevant cases only.

Assumption 2.3. *Assume that* $c_{q\theta}^2 > c_{\theta\theta}c_{qq}$.

This assumption implies that first best welfare is convex in θ. Hence, we only need to consider two cases. Either first best welfare is increasing in θ or it is first decreasing and then increasing in θ. Further, we can show that first best quality increases faster with θ than $k(\theta)$ and therefore k can intersect q^{fb} at (at most) one type; a result that we use below.

Lemma 2.1. *First best welfare* $W^{fb}(\theta)$ *is convex in* θ *and* $q_\theta^{fb}(\theta) > k_\theta(\theta)$.

Now we define the two cases that we focus on in this paper.

Definition 2.1. *We consider the two cases*

(WM) *where first best welfare is monotone in* θ*:* $\frac{dW^{fb}(\theta)}{d\theta} > 0$ *for all* $\theta \in [\underline{\theta}, \bar{\theta}]$ *or*

(WNM) *where a* θ_w *exists such that* $\frac{dW^{fb}(\theta)}{d\theta} < 0$ *for* $\theta \in [\underline{\theta}, \theta_w)$ *and* $\frac{dW^{fb}(\theta)}{d\theta} > 0$ *for* $\theta \in (\theta_w, \bar{\theta}]$*; further* $W^{fb}(\bar{\theta}) > W^{fb}(\underline{\theta})$.

Hence, we exclude the case where $W^{fb}(\bar{\theta}) < W^{fb}(\underline{\theta})$ (and by lemma 2.1 this is the only case we exclude). In words, we keep on thinking of high (enough) θ as better.[10]

The following two examples give cost and surplus functions that correspond to cases (WM) and (WNM) resp.

Example 2.1. *Assume* $S(q) = q$ *and* $c(q, \theta) = (q-\theta)^2 + q(1-\theta/2)$ *where* θ *is distributed uniformly on* $[0, 1]$. *With these functions* $k(\theta) = 4\theta/5$ *and* $q^{fb}(\theta) = 5\theta/4$. *First best welfare is* $W^{fb}(\theta) = \frac{9}{16}\theta^2$ *which is increasing in* $\theta \in [0, 1]$.

The interpretation of this example could be that by the qualification of its staff a firm has the "natural quality level" θ. Producing at different qualities involves adjustment costs that increase with the distance $|q - \theta|$. Additionally, there is a linear cost of quality, e.g. from additional (non-staff) input factors. A high type firm, e.g. a firm that traditionally has had highly qualified staff and therefore is experienced in high quality production, has lower additional costs of quality.

[10] It will become clear that the opposite case with $W^{fb}(\bar{\theta}) < W^{fb}(\underline{\theta})$ is symmetric and does not need to be considered separately.

Example 2.2. *Assume* $S(q) = Sq$ *and* $c(q,\theta) = \frac{1}{2}q^2 - \theta q + \theta k$ *with* $k \in (S+\underline{\theta}, S+\bar{\theta})$. *Thus* $k(\theta) = k$ *in assumption 2.2. Then we find that* $q^{fb}(\theta) = S+\theta$ *and* $dW^{fb}(\theta)/d\theta = S+\theta-k$. *Hence, with* $(k-S) \in (\underline{\theta}, \bar{\theta})$ *first best welfare increases for* $\theta > k - S$ *and decreases for* $\theta < k - S$.

The second example reflects the standard idea that a firm with high fixed costs ($k\theta$) has lower marginal costs ($c_q = q - \theta$) of producing quality. That is, a firm that produces with a more capital intensive technology might have lower marginal costs for quality but higher fixed costs.

Now we are able to set up the mechanism design problem. The planner only needs one firm to supply the desired service or product. Since $n \geq 2$ firms are able to supply, the planner needs to determine: which firm wins the procurement, what quality level should this firm supply and how much money should be transferred to firms in return for this.

Let $t(\theta)$ denote the (expected) transfer paid by the planner to a firm of type θ and $x(\theta)$ the probability that type θ wins the procurement. That is, the planner offers a menu of choices for firms and each firm chooses the option that maximizes its profits. The planner's objective is to maximize the expected value of $Sq - t$. The payoff for a type θ player that chooses option (q, x, t) is written as $t - xc(q,\theta)$.[11]

Following Myerson (1981), we use a direct revelation mechanism. That is, we design a menu of choices where $(q(\theta), x(\theta), t(\theta))$ is the choice "meant for" type θ. Then we make it incentive compatible (IC) for type θ to choose this option. That is, it is IC for θ to truthfully reveal his type.

Type θ can misrepresent as $\hat{\theta}$ and its profits equal

$$\pi(\hat{\theta}, \theta) = t(\hat{\theta}) - x(\hat{\theta})c(q(\hat{\theta}), \theta) \tag{2.3}$$

A menu $q(\cdot), x(\cdot), t(\cdot)$ is IC if and only if

$$\Phi(\hat{\theta}, \theta) \equiv \pi(\theta, \theta) - \pi(\hat{\theta}, \theta) \geq 0 \tag{2.4}$$

for all $\theta, \hat{\theta} \in [\underline{\theta}, \bar{\theta}]$.

[11] Note that since firms' and planner's utility is quasilinear in money, it is without loss of generality to assume that transfer payments t are paid without conditioning on winning: A price p which is paid only when winning the auction is equivalent to an unconditional transfer $t = px$.

With a slight abuse of notation we define the function $\pi(\theta)$ as:

$$\pi(\theta) = \max_{\hat{\theta}} \pi(\hat{\theta}, \theta)$$

Hence, using an envelope argument, incentive compatibility implies

$$\pi_\theta(\theta) = -x(\theta)c_\theta(q(\theta), \theta) \tag{2.5}$$

This equation makes sure that the first order condition for truthful revelation of θ is satisfied. The next result derives a tractable form for the local second order condition.

Lemma 2.2. *For the second order conditions to be locally satisfied, we also need that*

$$x_\theta(\theta)c_\theta(q(\theta), \theta) + x(\theta)c_{q\theta}(q(\theta), \theta)q_\theta(\theta) \leq 0. \tag{SOC}$$

As shown in textbooks like Fudenberg and Tirole (1991), first and second order conditions above imply global IC (as in equation (2.4)) if $c_\theta < 0$ for all $q \in Q$. Because we assume that firms are specialized (assumption 2.2), local IC does not automatically imply global IC. Hence, we need to verify explicitly below that global IC is satisfied.

Intuitively, assumption 2.2 is similar to a violation of single crossing. Viewing firm's payoff, $t - xc(q, \theta)$ as a function of x, the standard single crossing assumption would require that the derivative of $t - xc(q, \theta)$ with respect to x is monotone in type, i.e. single crossing would require that c_θ does not change sign. But assumption 2.2 states exactly the opposite. It is well known that in models without single crossing non-local IC can become relevant, see for example Araujo and Moreira (2010) or Schottmüller (2011a). We will first neglect these non-local incentive constraints and verify ex post that they do not bind. Although there are some issues with defining single crossing in multidimensional models (see for example McAfee and McMillan (1988)), we refer to c_θ switching sign as a "violation of single crossing".

Finally, because c_θ can switch sign, it is not clear that $\pi(\underline{\theta}) = 0$ under the optimal mechanism. That is, we cannot rule out that $\pi(\underline{\theta}) > 0$ while $\pi(\theta) = 0$ for some $\theta > \underline{\theta}$. Hence, we need to explicitly track the individual rationality constraint

$$\pi(\theta) \geq 0 \tag{2.6}$$

where we normalize firms' outside option to zero.

We assume that the planner maximizes utility Sq minus the transfer paid to firms. If the planner assigns the project to player i with probability x^i where i produces quality q^i and receives transfer t^i, the planner's utility from i can be written as $x^i Sq^i - t^i = x^i(Sq^i - c^i) - \pi^i$. Above, we did not index q and x by $i = 1 \dots n$ although we have n firms. It will be shown now that this is indeed unnecessary because of the symmetry of the problem. To do so, we write the planner's optimization problem[12] including the firm identifier i

$$\begin{aligned}
&\max_{q^i, x^i, \pi^i} \int_{\underline{\theta}}^{\bar{\theta}} \dots \int_{\underline{\theta}}^{\bar{\theta}} \sum_{i=1}^{n} f(\theta^1) \dots f(\theta^n)/f(\theta^i) \left\{ f(\theta^i)[x^i(\Theta)(Sq^i(\theta^i) - c(q^i(\theta^i), \theta^i)) - \pi^i(\Theta)] \right. \\
+ \quad & \lambda^i(\theta^i)(\pi^i_{\theta^i}(\Theta) + x^i(\Theta) c_{\theta^i}(q^i(\theta^i), \theta^i)) \\
- \quad & \mu^i(\theta^i)(x^i_{\theta^i}(\Theta) c_{\theta^i}(q^i(\theta^i), \theta^i) + x^i(\Theta) c_{q\theta^i}(q^i(\theta^i), \theta^i) q^i_{\theta^i}(\theta^i)) \\
+ \quad & \left. \eta^i(\theta)\pi^i(\Theta) \right\} - \sum_i \{\tau^i(\Theta) x^i(\Theta)\} + \sigma(\Theta) \left(1 - \sum_i x^i(\Theta) \right) \, d\theta_1 \dots d\theta_n
\end{aligned}$$

where $\lambda^i(\cdot)$ and $\mu^i(\cdot)$, $\eta^i(\cdot) \geq 0$ are the Lagrange multipliers (co-state variables) of the constraints (2.5), (SOC) and (2.6). Here, $x^i(\Theta)$ denotes the probability of firm i being contracted when types are $\Theta = (\theta^1 \dots \theta^n)$. The last constraint ensures that probabilities sum to no more than 1. Because of assumption 2.1, this constraint will bind and $\sigma(\Theta)$ will therefore be positive. The second but last term secures nonnegativity of the contracting probabilities where the Lagrange multiplier $\tau^i(\Theta) \geq 0$.

The Euler equation for $x^i(\Theta)$ can be rewritten as

$$\begin{aligned}
f(\theta^1) \dots f(\theta^n)/f(\theta^i) \left\{ f(\theta^i)[Sq^i(\theta^i) - c(q^i(\theta^i), \theta^i)] + \lambda^i(\theta^i) c_{\theta^i}(q^i(\theta^i), \theta^i) \right. \\
\left. + \mu^i_{\theta^i}(\theta^i) c_{\theta^i}(q^i(\theta^i), \theta^i) + \mu^i(\theta^i) c_{\theta^i\theta^i}(q^i(\theta^i), \theta^i) \right\} = \sigma(\Theta) + \tau^i(\Theta). \quad (2.7)
\end{aligned}$$

As the objective function is linear in $x^i(\cdot)$, we get what is called a "bang-bang" solution in optimal control theory: For any Θ, the firm i with the highest left hand side in (2.7) is contracted, i.e. $x^i(\Theta) = 1$, while the other firms are not, i.e. $x^j(\Theta) = 0$ for all $j \neq i$.

With this simple structure for the decision $x(\Theta)$, the maximization problem is totally symmetric across all i. In particular, the first order conditions for $q^i(\cdot)$ and $\pi^i(\cdot)$ are the same for all i. Consequently, we can use a notationally much simpler formulation of the

[12] We immediately focus on the case with non-random qualities, i.e. each type's quality is a deterministic function of his type only. Appendix E in Jullien (2000) can be easily adapted to our setting to show that optimal mechanisms are indeed deterministic under our assumptions.

maximization problem

$$\begin{aligned} & \max \int_{\underline{\theta}}^{\bar{\theta}} f(\theta)[x(\theta)(Sq(\theta) - c(q(\theta),\theta)) - \pi(\theta)] \\ + \quad & \lambda(\theta)(\pi_\theta(\theta) + x(\theta)c_\theta(q(\theta),\theta)) \\ - \quad & \mu(\theta)(x_\theta(\theta)c_\theta(q(\theta),\theta) + x(\theta)c_{q\theta}(q(\theta),\theta)q_\theta(\theta)) \\ + \quad & \eta(\theta)\pi(\theta)d\theta \end{aligned} \tag{2.8}$$

where $\lambda(\cdot)$ and $\mu(\cdot), \eta(\cdot) \geq 0$ are the Lagrange multipliers (co-state variables) of the constraints (2.5), (SOC) and (2.6) respectively.

The Euler equation for $\pi(\cdot)$ implies

$$\lambda_\theta(\theta) = -f(\theta) + \eta(\theta) \tag{2.9}$$

The first order condition for $q(\cdot)$ can be written as

$$f(\theta)(S - c_q(q(\theta),\theta)) + \lambda(\theta)c_{q\theta}(q(\theta),\theta) + \mu(\theta)c_{q\theta\theta}(q(\theta),\theta) = -\mu_\theta(\theta)c_{q\theta}(q(\theta),\theta). \tag{2.10}$$

Define the virtual valuation of type θ as

$$VV(\theta) = Sq(\theta) - c(q(\theta),\theta) + \frac{\lambda(\theta)}{f(\theta)}c_\theta(q(\theta),\theta). \tag{2.11}$$

If constraint (SOC) is not binding, the planner's objective function is linear in $x(\theta)$, where $x(\theta)$ is multiplied by $VV(\theta)$. Hence, using standard arguments, the firm with the highest VV wins the procurement contract. The virtual valuation includes next to the first best welfare a rent extraction term. Roughly speaking, contracting a type more often, i.e. increasing $x(\theta)$, changes the slope of the rent function $\pi(\theta)$; see equation (2.5). If, for example, the incentive constraints is downward binding and the rent function is increasing more steeply, types above θ will get a higher rent. $\lambda(\theta)$ is basically the weight of the types that benefit from such a change.

The following two lemmas are useful in the analysis below. The first lemma establishes that we have a monotone hazard rate property for our case with specialized firms.

Lemma 2.3. *If either*

(i) $\lambda(\theta) \geq 0$ *for all* $\theta \in [\underline{\theta}, \bar{\theta}]$ *and* $\lambda(\bar{\theta}) = 0$ *or*

(ii) $\lambda(\underline{\theta}) = \lambda(\bar{\theta}) = 0$,

then

$$\frac{d(\lambda(\theta)/f(\theta))}{d\theta} < 0 \tag{2.12}$$

for values of θ *with* $\eta(\theta) = 0$.

As we will see below, the property in equation (2.12) is useful to have. It is part of the set of conditions to make quality q monotone in θ. Case (i) is relevant for the (WM) case and case (ii) for (WNM). If $\eta(\theta) > 0$, it turns out that the monotonicity of quality is easy to prove (see the discussion of profit bunching below).

Lemma 2.4. *Assume* $\mu(\theta) = 0$ *and* $\frac{d(\lambda(\theta)/f(\theta))}{d\theta} < 0$ *for all* θ *with* $\eta(\theta) = 0$. *Then*

1. *if there is* $\hat{\theta}$ *such that* $q(\hat{\theta}) = k(\hat{\theta})$ *then* $q_\theta(\hat{\theta}) \geq k_\theta(\hat{\theta})$,
2. *if there is* θ' *such that* $c_\theta(q(\theta'), \theta') \leq 0$ *then* $c_\theta(q(\theta), \theta) \leq 0$ *for all* $\theta > \theta'$ *and*
3. *if there exist* $\theta_1, \theta_2 > \theta_1$ *with* $\pi(\theta_1) = \pi(\theta_2) = 0$ *then* $\pi(\theta) = 0$ *for all* $\theta \in [\theta_1, \theta_2]$.

The first result says that if q and k coincide for some value $\hat{\theta}$, then it cannot be the case that k exceeds q for higher values of θ. Further, it is the case that once $c_\theta \leq 0$ for the optimal $q(\theta)$ then c_θ stays non-positive for all higher θ. Finally, the third result implies that if two types have zero profits then all types in between have zero profits as well. That is, there cannot be a type $\theta \in [\theta_1, \theta_2]$ with positive profits (and negative profits are excluded by equation (2.6)).

Finally, we use the following notation. Let $q^h(\theta)$ denote the solution to[13]

$$S - c_q(q(\theta), \theta) + \frac{1 - F(\theta)}{f(\theta)} c_{q\theta}(q(\theta), \theta) = 0 \tag{2.13}$$

and $q^l(\theta)$ the solution to[14]

$$S - c_q(q(\theta), \theta) - \frac{F(\theta)}{f(\theta)} c_{q\theta}(q(\theta), \theta) = 0 \tag{2.14}$$

In the following two sections we solve the problem for the WM and then the WNM case. The strategy will be to solve the first order condition and then to verify ex post that (SOC) and non-local incentive constraints do not bind under our assumptions. Section 2.5 returns to the case where (SOC) is not satisfied.

[13] If several q solve this equation, we denote the highest by q^h. By assumption 2.1 and 2.2, there can be at most one $q > k(\theta)$ satisfying equation (2.13).

[14] If the solution to this equation is not unique, let the lowest solution be q^l. By assumption 2.1 and 2.2, there is at most one $q < k(\theta)$ satisfying equation (2.14).

2.3. First best welfare monotone

We will now characterize the optimal mechanism for the WM-case. The following lemma is useful to characterize the optimal menu. The lowest type $\underline{\theta}$ receives lowest profits (zero) and the IC constraint (2.5) is binding downwards. That is, high types would like to mimic low types (not the other way around).

Lemma 2.5. *In the WM-case we have:* $\pi(\underline{\theta}) = 0$ *and* $\lambda(\theta) \geq 0$ *for all* $\theta \in [\underline{\theta}, \bar{\theta}]$.

Now we are able to characterize the solution for the WM case. There are two cases to consider. In the first case, the solution (given by equation (2.13)) is such that the specialization of the firms plays no role. This is basically the solution to a standard problem. In the second case, low types up to a type θ_b are bunched on zero profits (but with different quality levels) and from $\theta_b \geq \underline{\theta}$ onwards, $q(\theta)$ follows the solution in equation (2.13).

Proposition 2.1. *There are two cases:*

1. *If* $c_\theta(q^h(\underline{\theta}), \underline{\theta}) < 0$, *then* $q^h(\theta)$ *in equation (2.13) gives the optimal quality for all* $\theta \in [\underline{\theta}, \bar{\theta}]$. *We have* $\pi_\theta(\theta), q_\theta(\theta), x_\theta(\theta) > 0$ *for each* $\theta \in [\underline{\theta}, \bar{\theta}]$.

2. *If* $c_\theta(q^h(\underline{\theta}), \underline{\theta}) \geq 0$ *then there exists a largest* $\theta_b \geq \underline{\theta}$ *such that*

$$q(\theta) = k(\theta) \text{ for all } \theta \in [\underline{\theta}, \theta_b]$$

and θ_b *is determined by the unique solution to*

$$S - c_q(k(\theta_b), \theta_b) + \frac{1 - F(\theta_b)}{f(\theta_b)} c_{q\theta}(k(\theta_b), \theta_b) = 0 \tag{2.15}$$

For all $\theta > \theta_b$ *quality* $q(\theta) = q^h(\theta)$. *We have*

$$\pi(\theta) = 0 \text{ for all } \theta \in [\underline{\theta}, \theta_b],$$
$$\pi_\theta(\theta) > 0 \text{ for all } \theta \in (\theta_b, \bar{\theta}], \text{ and}$$
$$x_\theta(\theta), q_\theta(\theta) \geq 0 \text{ for all } \theta \in [\underline{\theta}, \bar{\theta}].$$

The relaxed solution is globally incentive compatible.

In the first case of proposition 2.1, the possibility that c_θ can change sign does not play a role in the relevant range of q. Therefore, the standard menu as in Che (1993) results. In the second case, c_θ would be positive for some types in the standard quality menu which is given by (2.13). A direct corollary of lemma 2.1 is that $c_\theta \leq 0$ at the first best quality level. Hence, the standard downward distortion of q caused by the rent extraction motive is responsible for having $c_\theta > 0$ for some types under q^h. By (2.5), profits are decreasing at types where $c_\theta > 0$. If q^h was implemented, type θ^b would therefore have zero profits while lower types would have positive profits. But now the principal can do better than q^h: By assigning $k(\theta)$ to types below θ^b, the principal (i) saves rents as those types remain at zero profits and (ii) reduces distortion compared to q^h. Because each type is most cost efficient at his $k(\theta)$, no other type can profitably misrepresent as θ if θ expects zero profits and produces quality $k(\theta)$. Put differently, the incentive constraint is lax in this situation. Therefore, it is not necessary to distort quality further down than $k(\theta)$ for rent extraction purposes. In some sense, specialization leads to "less distortion at the bottom" and more rent extraction.

In conclusion, the menu in case 2 of proposition 2.1 consists of a standard part for high types and one part where types produce at $k(\theta)$ and consequently the incentive constraint is lax.

2.4. First best welfare non-monotone

In this section, first best welfare is U-shaped. The lowest type $\underline{\theta}$ is no longer worst (in a first best sense) and therefore he might have positive profits under the optimal mechanism. The following lemma confirms this intuition.

Lemma 2.6. *Under WNM,* $\pi(\underline{\theta}) > 0$, $\pi(\bar{\theta}) > 0$ *and* $\lambda(\underline{\theta}) = \lambda(\bar{\theta}) = 0$.

One can think of the WNM case as having two standard menus. One for lower θ in which lower types are better, profits are decreasing in type and quality is distorted upwards. The other for higher θ with higher types being better, profits increasing in type and quality distorted downwards. These two menus have to be reconciled.

In the principal agent literature, irregularities are often dealt with bunching types on

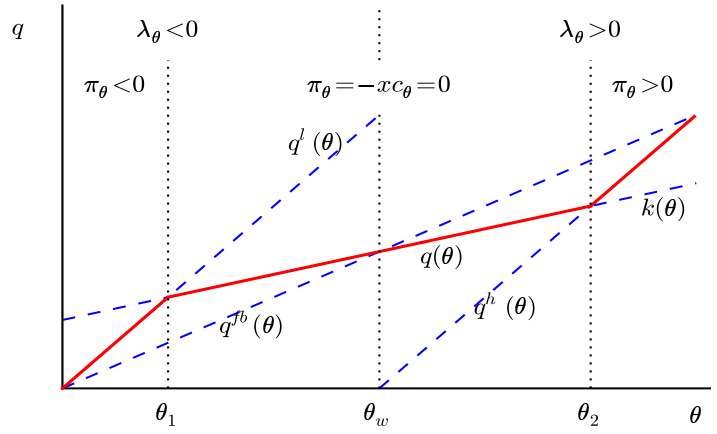


Figure 2.1: Optimal $q(\theta)$ (solid, red) in the WNM case, together with (dashed) $q^l(\theta), q^{fb}(\theta), k(\theta), q^h(\theta)$.

one decision[15]. Hence, a first idea could be that bunching on quality might be used to connect the two menus. It is quickly shown that this does not work. To see this, suppose –by contradiction– that $q(\theta) = q^b$ for types θ in the bunching interval. As profits are decreasing in θ for low θ and increasing in θ for high θ, the type θ' with the lowest profits ($\pi(\theta') = 0$) would have to be in the bunching interval. From (2.5), the profit minimizing type has to satisfy $c_\theta(q^b, \theta') = 0$. Hence, he produces at $q^b = k(\theta')$ and is for this quality level the most efficient type. But then he has the highest profits of all types in the quality-bunching interval. This contradiction implies that a menu with quality bunching cannot be the solution.[16]

The right way to reconcile the two standard menus is an interval of types with zero profits (but differing quality levels). Incentive compatibility within the bunched interval is no problem here. Each bunched type θ will produce at quality level $k(\theta)$ at which he has lower costs than any other type. The following proposition describes the optimal menu in the WNM case.

Proposition 2.2. *There exist unique* θ_1 *and* θ_2*, with* $\theta_1 < \theta_2$*, such that* $q^l(\theta_1) = k(\theta_1)$

[15] See, for instance, Guesnerie and Laffont (1984) or Fudenberg and Tirole (1991, ch. 7).

[16] Unless it happens at $q^b = k$ in the case where $k(\theta) = k$ is constant (on a subset of $[\underline{\theta}, \bar{\theta}]$). In this case, those types that are bunched on zero profits would also have the same quality $q(\theta) = k$.

and $q^h(\theta_2) = k(\theta_2)$*. Quality is determined by*

$$q(\theta) = \begin{cases} q^h(\theta) & \text{for all } \theta > \theta_2 \\ k(\theta) & \text{for all } \theta \in [\theta_1, \theta_2] \\ q^l(\theta) & \text{for all } \theta < \theta_1. \end{cases} \tag{2.16}$$

We have

$$\begin{aligned} \pi(\theta) &= 0 \text{ for all } \theta \in [\theta_1, \theta_2] \\ \pi_\theta(\theta) &< 0 \text{ for all } \theta < \theta_1 \\ \pi_\theta(\theta) &> 0 \text{ for all } \theta > \theta_2 \\ q_\theta(\theta) &\geq 0. \end{aligned}$$

Type θ_w*, who has the lowest first best welfare of all types, is in the zero profit interval and produces his first best quality. It holds that*

$$x_\theta(\theta) < 0 \text{ for all } \theta < \theta_w$$
$$x_\theta(\theta) > 0 \text{ for all } \theta > \theta_w.$$

The relaxed decision is globally incentive compatible.

Figure 2.1 illustrates proposition 2.2. Quality is above first best, i.e. upwards distorted, for low θ and downwards distorted for high θ. This is a consequence of the U-shaped first best welfare which implies that low types are better around $\underline{\theta}$ and high types are better around $\bar{\theta}$. Quality is not distorted at the (locally) best types $\underline{\theta}$ and $\bar{\theta}$ which resembles the well known "no distortion at the top" result. Quality is also undistorted for the worst type θ_w which allows a continuous transition from upwards to downwards distortion.

The boundaries of the zero profit interval $[\theta_1, \theta_2]$ are at those types where the low standard menu and the high standard menu feature $q(\theta) = k(\theta)$. In the zero profit interval, each type produces the quality for which he is the cost minimizing type, i.e. $k(\theta)$. Any other quality could not be incentive compatible within a zero profit interval as either types slightly higher or slightly lower would be more efficient. But then they could achieve positive profits by misrepresenting. From $q(\theta) = k(\theta)$, it is evident that misrepresenting as any other type $\theta \in [\theta_1, \theta_2]$ cannot be profitable and this is exactly the reason why the zero profit types do not receive any informational rent.

At θ_1 and θ_2, $q(\theta)$ is kinked. At θ_1, for example, the quality according to the standard low menu (q^l) would include additional informational distortion pushing quality upwards. Therefore $q^l(\theta) > k(\theta)$ for types slightly above θ_1 while $q(\theta) = k(\theta)$ is necessary to stay in the zero profit interval.

Note that for types above θ_w the optimal contract is similar to the one derived in proposition 2.1, i.e. quality and virtual valuation are the same. This is quite intuitive as first best welfare is increasing for those types. In this sense, proposition 2.2 "extends" proposition 2.1.

The following proposition formalizes the "grudge" of high θ incumbents against low θ entrants: although in second best the incumbent generates higher quality and higher welfare than the entrant, it can happen that the entrant wins the procurement contract. Incidentally, the opposite can happen as well: an incumbent wins from an entrant who generates higher (second best) welfare.

Proposition 2.3. *The optimal allocation is not (second best) efficient in the sense that there exist types θ', θ'' such that θ' wins against θ'' while $W^{sb}(\theta'') > W^{sb}(\theta')$.*[17]

A similar result is well known in auctions with asymmetric bidders. Myerson (1981) shows that it is optimal to discriminate between bidders drawing their valuations from different distributions. For example, if bidder A draws his valuation from a distribution putting more weight on high values and bidder B draws from a distribution with low values, the auction will favor B. This decreases the rents A will get by stimulating him to bid more aggressively. In our case, there is only one distribution from which types are drawn. Nevertheless, the intuition is similar. The reason for discrimination are informational distortions. For the lower standard menu, the relevant term inducing distortion in the virtual valuation is $-F(\cdot)c_\theta(\cdot)$. For high θ, the respective term is $(1-F(\cdot))c_\theta(\cdot)$. While discrimination in Myerson (1981) results from the fact that different distributions govern the distortion, discrimination in our model is due to different parts of the same distribution governing distortion: For low θ, the left tail is relevant and for high types the right tail of the distribution matters for distortion. The reason is that the local incentive constraint is upward binding in the lower standard menu and

[17] We use the term *second best efficient* to describe a situation where the selection rule picks the firm providing the highest W^{sb}. W^{sb} is welfare under the optimal quality schedule derived in propositions 2.1 and 2.2.

downward binding in the upper standard menu. On a more intuitive level, by ex ante committing to let a worse low type $\theta' < \theta^w$ win against a better high type $\theta'' > \theta^w$, one can save informational rents for θ'' and all types above him. The reason is that the probability that θ'' wins the auction, i.e. $x(\theta'')$, decreases and therefore the slope of the rent function $\pi_\theta(\theta'') = x(\theta'')c_\theta(q(\theta''), \theta'')$ decreases. Loosely speaking, one stimulates θ'' and higher types to bid more aggressively.

We conclude this section with a brief discussion of how to implement the optimal menus in propositions 2.1 and 2.2. We argue that this is more straightforward for the WM than for the WNM case. In each case, we have in mind that the government announces at the start its willingness p to pay (conditional on winning) for each quality level q. Hence, p corresponds to t/x in the mechanism design notation used so far. In case 1 in proposition 2.1, the government can then organize a second price auction to determine the firm that wins the contract. The firm with the highest bid, wins and pays the second highest bid. This firm is then allowed to choose its combination (q, p) from the menu announced by the government. Since the planner wants the highest type to win and profits are strictly increasing in θ, such an auction selects the right type as winner. Since we assume that the service is valuable enough that it has to be supplied, there is no reserve price in this auction.

However, in the second case in proposition 2.1 there are a number of types with equal (zero) profits while the planner prefers higher θ for the case where quality increases in θ. The auction described above is not optimal here since it cannot discriminate between types with the same profits. In that case, the selection mechanism must be based on quality directly. To be more precise, let firms bid qualities. The firm bidding the highest quality wins, produces this quality and receives payment p (according to the menu announced by the government).

In the WNM case, the auctions described above do not implement the optimal mechanism. The planner's preference over winning types is given by the virtual valuation VV in equation (2.11). Again, a price auction cannot work because of the zero profit interval: These types have the same valuation for winning the auction but should have different probabilities of winning ($VV(\theta)$ and thus $x(\theta)$ is not constant over $\theta \in [\theta_1, \theta_2]$). An auction based on quality does not work either because q_θ and x_θ do not have the same sign for all types. Furthermore, a scoring rule auction, as analyzed in Che (1993),

cannot implement the optimal mechanism either. This can be seen as follows.

In a scoring rule auction, each bidder bids a score and the bidder with the highest score is contracted. Consider a scoring rule of the form $score(p,q) = s(q) - p$, where the price p is (only) paid to the winner of the auction. In a second score auction, the winner has to provide (q,p) which corresponds to the second highest score. Hence, the second highest score, $score^{(2)}$, determines the rents going to the winner. Consequently, it is a dominant strategy to bid the maximum score that one can deliver at non-negative profits. Thus,

$$bid(\theta) = \max_q \{s(q) - c(q,\theta)\}.$$

To implement the optimal mechanism, it must be the case that the first order condition

$$s_q(q) - c_q(q,\theta) = 0$$

yields $q(\theta)$ as given by equation (2.16). As shown by Che (1993), it then follows that

$$s(q) = Sq + \int_{q(\underline{\theta})}^{q} \frac{\lambda(q^{-1}(s))}{f(q^{-1}(s))} c_{q\theta}(s, q^{-1}(s))\, ds \qquad \text{for} \qquad q \in [q(\underline{\theta}), q(\bar{\theta})]$$

and $-\infty$ for all other q; where $q^{-1}(s)$ is the inverse of $q(\theta)$ and $\lambda(\theta)$ is the Lagrange multiplier (co-state) of the optimal menu derived in proposition 2.2.

This implies that the winner is determined by the firm bidding the highest value of

$$bid(\theta) = Sq - c(q,\theta) + \int_{q(\underline{\theta})}^{q} \frac{\lambda(q^{-1}(s))}{f(q^{-1}(s))} c_{q\theta}(s, q^{-1}(s))\, ds$$

while in the optimal mechanism, the winner has the highest value of VV as given by equation (2.11).[18] Put differently, if the scoring rule implements the optimal mechanism it has to hold that $bid(\theta') = bid(\theta'')$ whenever $VV(\theta') = VV(\theta'')$ under the optimal mechanism. The following proposition says that generically this is untrue under WNM.

Proposition 2.4. *Generically, a simple scoring rule auction cannot implement the optimal mechanism in the WNM case.*

Consequently, more general mechanisms are needed for implementation in the WNM case. As shown in proposition 2.3, the (optimal) government's decision may be criticized

[18] Note that there is also an issue in choosing the right tie-breaking rule. From the envelope theorem, $\frac{d\,bid(\theta)}{d\theta} = -c_\theta(q(\theta),\theta)$. Therefore, all types with zero profits have the same bid because $q(\theta) = k(\theta)$ for them which implies $\frac{d\,bid(\theta)}{d\theta} = 0$. The tie breaking rule should follow $VV(\theta)$ here, making the mechanism not easy to implement.

ex post in case a firm loses from a winner generating lower (second best) welfare. If the government cannot implement the optimal mechanism because of its complexity, more inefficiencies will be introduced in the WNM case.

2.5. Robustness

Above we made some assumptions on third derivatives of the cost function and the distribution of θ for ease of exposition. Here we discuss how the solution changes when these assumptions are no longer satisfied. In principle, there are two possible problems that can arise: First, the second order condition (SOC) could be violated in the derived solution. Second, the program is no longer globally concave.

2.5.1. Violation second order condition

For concreteness, we focus here on the WM case and assume that the problems arise because of a violation of the MHR assumption. The cases where third derivatives cause problems with (SOC) are dealt with analogously. In the WM case, the change in q for $\theta > \theta_b$ is given by

$$q_\theta(\theta) = \frac{c_{q\theta}(q(\theta),\theta) - c_{q\theta\theta}(q(\theta),\theta)\frac{1-F(\theta)}{f(\theta)} - c_{q\theta}(q(\theta),\theta)\frac{d\left(\frac{1-F(\theta)}{f(\theta)}\right)}{d\theta}}{-c_{qq}(q(\theta),\theta) + c_{qq\theta}(q(\theta),\theta)\frac{1-F(\theta)}{f(\theta)}}. \tag{2.17}$$

The assumptions made above are sufficient conditions for $q_\theta(\theta) \geq 0$. Hence, if F does not satisfy the MHR assumption, it can still be the case that $q_\theta(\theta) \geq 0$ and $x_\theta(\theta) \geq 0$.[19] If q and x are non-decreasing in θ, we know that the second order condition (SOC) is satisfied. Even if, say, $q_\theta(\theta) < 0$ while $x_\theta(\theta) > 0$, equation (SOC) can still be satisfied.

Now we consider the case where $d((1 - F(\theta))/f(\theta))/d\theta > 0$ for $\theta > \theta_b$ in such a way that $q_\theta < 0$ causes a violation of (SOC). We first sketch how this is dealt with in general. Then we work out an example. As shown by Guesnerie and Laffont (1984) and Fudenberg and Tirole (1991) for the case of a single dimensional decision (say, only quality), a violation of the second order condition leads to bunching: several θ-types produce the same quality. However, in our case the decision is two dimensional: quality q and the probability of winning x. In fact, below we do not work with x but with the

[19] Whether $x_\theta \geq 0$ can be derived from the expression for $dVV(\theta)/d\theta$ in equation (2.37).

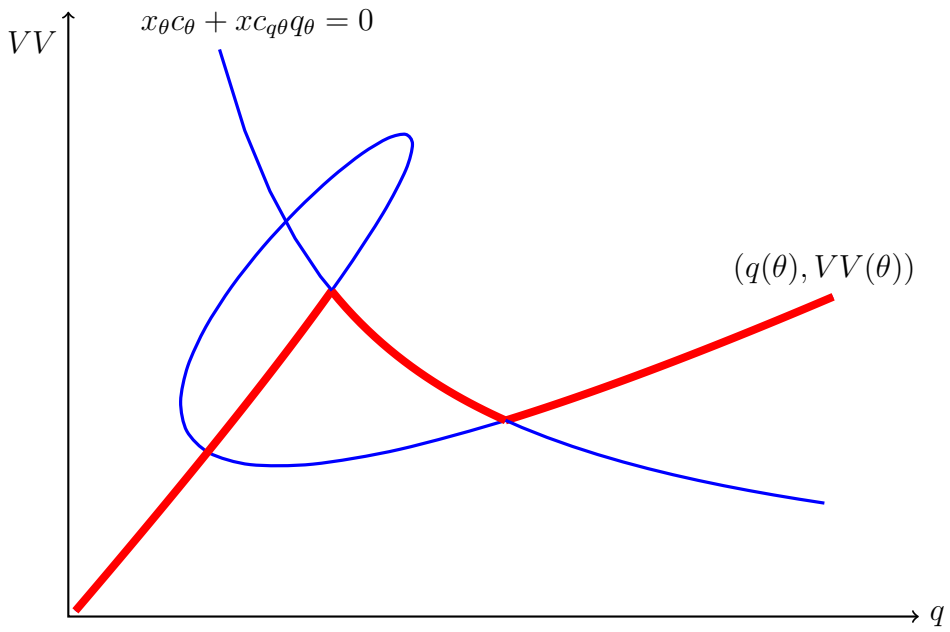


Figure 2.2: Solution for quality $q(\theta)$ and virtual valuation $VV(\theta)$ for the case where (second order) condition (SOC) is violated.

virtual valuation VV as there is a one-to-one relation between the two (i.e. higher VV implies higher x and the other way around). We show that in this two-dimensional case, it is not necessarily true that a violation of (SOC) leads to bunching of types θ to the same quality q and probability of winning x.[20]

We use figure 2.2 to illustrate the procedure. This figure shows equation (SOC) (where it holds with equality) in (q, VV) space and the solution $(q(\theta), VV(\theta))$ that follows from the planner's optimization problem while ignoring the second order condition; i.e. assuming $\mu_\theta(\theta) = 0$ for all θ. The former curve is downward sloping in the WM case since

$$\frac{dx}{dq} = \frac{x_\theta(\theta)}{q_\theta(\theta)} = -x(\theta)\frac{c_{q\theta}(q(\theta), \theta)}{c_\theta(q(\theta), \theta)} < 0$$

In the simple case (that we also use in the example below) where $c_{\theta\theta} = 0$, this curve boils down to

$$x(\theta)c_\theta(q(\theta), \theta) = -K < 0 \tag{2.18}$$

for some constant $K > 0$, as differentiating equation (2.18) with respect to θ indeed gives

[20] A related point is already made by García (2005). He shows in a multidimensional screening model where single crossing holds in all dimensions that non-monotone decisions can be optimal (even if second order conditions do not bind).

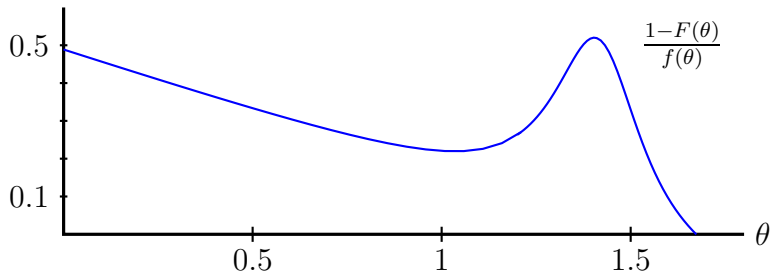


Figure 2.3: Inverse hazard rate with $f(\theta) = (\theta - a)^2 + 1/50$

$x_\theta c_\theta + x c_{q\theta} q_\theta = 0$.

The solution $(q(\theta), VV(\theta))$ ignoring the second order constraint, starts at $\underline{\theta}$ in the bottom left corner and moves first over the thick (red) part of this curve, then follows the thin (blue) part, curving back (i.e. both q and x fall with θ) then both q and x increase again with θ and we end on the thick (red) part of the curve. The part of the curve where $q_\theta, x_\theta < 0$ violates equation (SOC).

Hence, we need to find θ_a, θ_b where (SOC) starts to bind and $\mu(\theta) > 0$. Then from θ_a onwards, we follow the binding constraint till we arrive at θ_b, from which point onwards we follow the solution $(q(\theta), VV(\theta))$ again. As shown in figure, the choice of θ_a determines both the trajectory $(\tilde{q}(\theta), \tilde{VV}(\theta))$ satisfying equation (SOC) and the end point of this trajectory θ_b. Since $\mu(\theta) = 0$ both for $\theta < \theta_a$ and for $\theta > \theta_b$, it must be the case that $\int_{\theta_a}^{\theta_b} \mu_\theta(\theta) d\theta = 0$. To illustrate, for the case where $c_{q\theta\theta} = 0$,[21] this can be written as (using equation (2.10)):

$$\int_{\theta_a}^{\theta_b} \frac{f(\theta)(S - c_q(\tilde{q}(\theta), \theta)) + (1 - F(\theta) c_{q\theta}(\tilde{q}(\theta), \theta)}{c_{q\theta}(\tilde{q}(\theta), \theta)} d\theta = 0 \tag{2.19}$$

We now illustrate this approach with an example.

Example 2.3. *To violate the monotone hazard rate assumption we use the density* $f(\theta) = (\theta - a)^2 + 1/50$ *with support* $[0, a + 1/4]$ *where* a *has to be approximately* 1.42 *to satisfy the requirements of a probability distribution. The hazard rate of this distribution is depicted in figure 2.3.*

Assume that there are two firms and that $c(q, \theta) = \frac{1}{2}q^2 - q\theta + \theta$. *Then* $c_\theta(q, \theta) = 1 - q$ *which changes sign at* $q = 1$. *As* $c_{\theta\theta} = 0$, *the binding second order condition takes the*

[21] If $c_{q\theta\theta} \neq 0$, the differential equation (2.10) has to be solved for $\mu(\theta)$. Although a bit tedious, this is do-able since the differential equation is linear and first order in $\mu(\theta)$.

form of (2.18):

$$x = \frac{K}{q-1}$$

for some $K > 0$*. Note that this equation does not depend on* θ*. Hence, in this case, "following the constraint" takes the form of bunching* $\theta \in [\theta_a, \theta_b]$ *on some point*

$$(\tilde{q}, \tilde{VV}) \tag{2.20}$$

where $\tilde{VV}$ *corresponds to the probability* $\tilde{x} = \frac{K}{\tilde{q}-1}$*. Choosing* θ_a*, fixes* $\tilde{q} = q(\theta_a)$ *and* θ_b *since* $q(\theta_b) = \tilde{q}$*. Writing the dependency of* $\tilde{q}, \theta_b$ *on* θ_a *explicitly,* θ_a *solves equation (2.19):*

$$\int_{\theta_a}^{\theta_b(\theta_a)} f(\theta)(S - (\tilde{q}(\theta_a) - \theta)) - (1 - F(\theta))d\theta = 0 \tag{2.21}$$

Since equation (SOC) will already start to bind for θ_a *where* $q_\theta(\theta_a) > 0$*, it is routine to verify that this equation is downward sloping in* θ_a*. The unique solution in this example is* $\theta_a \approx 1.1685$ *which gives a corresponding* $\theta_b = 1.428$ *and* $\tilde{q} = 1.923$*.*

While the ironing procedure described above takes care of the local second order condition (SOC), this does not necessarily imply global incentive compatibility. Global constraints are mathematically intractable in general frameworks; see Araujo and Moreira (2010) and Schottmüller (2011a) for special examples of how to handle global constraints. However, the following proposition establishes that global constraints do not bind for a family of cost functions. This family includes the functions we used in the example and the most commonly used linear-quadratic cost functions.

Proposition 2.5. *If* $c_{\theta\theta} = 0$ *and the local second order condition (SOC) is satisfied, the solution is globally incentive compatible.*

2.5.2. Concavity in q

The second possible problem with $c_{qq\theta}$ not satisfying assumption 2.2 is that the planner's objective function (2.8) is not necessarily globally concave in $q(\cdot)$. However, in principle, the solution will still satisfy the first order conditions derived before. In particular, it is never optimal to choose $q \to \infty$: Since costs are convex and the principal's utility is linear in q, costs are higher than benefits for q high enough and therefore optimal qualities cannot be arbitrarily high. Hence, if Q is not bounded, the solution will be interior and satisfy the conditions derived above.

If the set of available qualities is a compact subset of $\mathbb{R}_+$, corner solutions could play a role; e.g. if quality cannot be higher than some level $\bar{q}$, some types might have $q(\theta) = \bar{q}$ and the first order conditions do not apply for them. However, such a situation can be easily approximated by a continuous cost function which is very steep around $\bar{q}$ (instead of jumping discontinuously to infinity) and to which our analysis would apply.

2.6. Conclusion

We analyzed a procurement setting in which the procurement agency cares not only about the price but also about the quality of the product. In many post liberalization situations incumbents seem to be good at producing high quality while entrants can produce low quality at very low costs. A similar pattern emerges if there are gains from specialization and firms can specialize in either high quality or low costs.

Standard procurement models do not account for this possibility because single crossing is assumed in all dimensions. More precisely, it is assumed that "type" denotes efficiency and not specialization. This implies that a more efficient type is simply better for all quality levels. We relax this assumption and allow each type to be specialized, i.e. to be the most most efficient type for some quality level. This leads to a bunching of types on zero profits. The intuition is that distorting quality further than the quality level a type is specialized in (for rent extraction reasons) is not necessary: A type producing "his quality level" with expected profits of zero cannot be mimicked by any other type. Hence, the incentive constraint is lax and an interval of zero profit types is feasible. In short, specialization limits distortion and helps the principal to extract rents.

If we assume that first best welfare is U-shaped, e.g. there are gains from specializing in low costs even from a welfare point of view, we get an interesting discrimination result. Types with lower second best welfare can be preferred to types with higher second best welfare. This is similar to auctions with asymmetric bidders where discriminatory mechanisms are well known. Contrary to this literature, bidders are drawn from the same distribution in our model. The intuition is that the incentive constraint is first upward and then, for higher types, downward binding. Therefore, different parts of the distribution govern the distortion for low and high types. By committing to let some

worse types win against some better types, the principal can reduce the rents of the best types. Loosely speaking, the better types are incentivized to bid more aggressively. Put differently, competitive pressure can be exerted even by firms that are clearly worse. Further, in this case "gold plating" can be optimal in the sense that some types produce quality levels above their first best levels.

2.7. Appendix: Proofs

Proof of lemma 2.1 From the first order condition for q^{fb} we derive that

$$q_\theta^{fb} = \frac{-c_{q\theta}}{c_{qq}} > 0$$

Then we find that

$$W_{\theta\theta}^{fb}(\theta) = \frac{c_{q\theta}^2}{c_{qq}} - c_{\theta\theta} > 0$$

from the assumptions made on the function $c(q, \theta)$. Further, it follows from $c_\theta(k(\theta), \theta) \equiv 0$ that $c_{q\theta}k_\theta(\theta) + c_{\theta\theta} = 0$. Hence, $q_\theta^{fb} > k_\theta$ if and only if

$$\frac{-c_{q\theta}}{c_{qq}} > \frac{c_{\theta\theta}}{-c_{q\theta}}$$

which holds by assumption 2.3. *Q.E.D.*

Proof of lemma 2.2 Define the function

$$\Phi(\hat{\theta}, \theta) = \pi(\theta, \theta) - \pi(\hat{\theta}, \theta) \geq 0$$

By IC this function is always positive and equal to zero if $\hat{\theta} = \theta$. In other words, the function Φ reaches a minimum at $\hat{\theta} = \theta$. Thus truth-telling implies both

$$\left.\frac{\partial \Phi(\hat{\theta}, \theta)}{\partial \hat{\theta}}\right|_{\hat{\theta}=\theta} = 0 \tag{2.22}$$

and

$$\left.\frac{\partial^2 \Phi(\hat{\theta}, \theta)}{\partial \hat{\theta}^2}\right|_{\hat{\theta}=\theta} \geq 0 \tag{2.23}$$

Since equation (2.22) has to hold for all $\hat{\theta} = \theta$, differentiating with respect to θ gives

$$\left.\frac{\partial^2 \Phi(\hat{\theta}, \theta)}{\partial \hat{\theta}^2}\right|_{\hat{\theta}=\theta} + \left.\frac{\partial^2 \Phi(\hat{\theta}, \theta)}{\partial \hat{\theta} \partial \theta}\right|_{\hat{\theta}=\theta} = 0$$

Then equation (2.23) implies that

$$\left.\frac{\partial^2 \Phi(\hat{\theta}, \theta)}{\partial \hat{\theta} \partial \theta}\right|_{\hat{\theta}=\theta} \leq 0$$

It follows from the definition of Φ that

$$\left.\frac{\partial^2 \Phi(\hat{\theta}, \theta)}{\partial \hat{\theta} \partial \theta}\right|_{\hat{\theta}=\theta} = x_\theta(\theta)c_\theta(q(\theta), \theta) + x(\theta)c_{q\theta}(q(\theta), \theta)q_\theta(\theta) \leq 0$$

which is the inequality in the lemma. *Q.E.D.*

Proof of lemma 2.3 We need to show that

$$\frac{d(\lambda(\theta)/f(\theta))}{d\theta} = \frac{\lambda_\theta(\theta)f(\theta) - \lambda(\theta)f_\theta(\theta)}{f(\theta)^2} < 0 \tag{2.24}$$

We consider the following four cases:

		$\lambda(\theta)$	
		≥ 0	< 0
$f_\theta(\theta)$	≥ 0	(α)	(β)
	< 0	(δ)	(γ)

Let's consider the two cases in the lemma in turn.

Case (i): We can solve

$$\lambda(\theta) = 1 - F(\theta) - \int_\theta^{\bar{\theta}} \eta(t)dt \tag{2.25}$$

Hence, we need to show

$$-f(\theta)^2 - \lambda(\theta)f_\theta(\theta) < 0 \tag{2.26}$$

where we use $\eta(\theta) = 0$. This is obviously satisfied in case (α). In case (δ) we have

$$-f(\theta)^2 - (1 - F(\theta) - \int_\theta^{\bar{\theta}} \eta(t)dt)f_\theta(\theta) < 0$$

Then this inequality is implied by the MHR assumption 2.1 where we write $d((1 - F(\theta))/f(\theta))/d\theta < 0$ as

$$-f(\theta)^2 - (1 - F(\theta))f_\theta(\theta) < 0 \tag{2.27}$$

because $\eta(t) \geq 0$. As we assume $\lambda(\theta) \geq 0$, we do not need to consider cases (β, γ).

Case (ii): Here we have a second way in which we can write $\lambda(\theta)$:

$$\lambda(\theta) = -F(\theta) + \int_{\underline{\theta}}^{\theta} \eta(t)dt \tag{2.28}$$

Equation (2.26) is clearly satisfied in cases $(\alpha), (\gamma)$. Case (δ) is satisfied for the same reason as above. Hence, we only need to consider case (β). Using equation (2.28), we write inequality (2.26) as

$$-f(\theta)^2 - (-F(\theta) + \int_{\underline{\theta}}^{\theta} \eta(t)dt)f_\theta(\theta) < 0$$

where we use $\eta(\theta) = 0$. Then this inequality is implied by the MHR assumption 2.1 where we write $d(F(\theta)/f(\theta))/d\theta > 0$ as

$$-f(\theta)^2 + F(\theta)f_\theta(\theta) < 0 \tag{2.29}$$

and $\eta(t) \geq 0$. *Q.E.D.*

Proof of lemma 2.4 We prove the parts in turn.

Part 1.: Suppose not, that is assume that $q(\hat{\theta}) = k(\hat{\theta})$ (i.e. $c_\theta(q(\hat{\theta}), \hat{\theta}) = 0$) and $q_\theta(\hat{\theta}) < k_\theta(\hat{\theta})$. Then for $\varepsilon > 0$ small enough, it is the case that

$$c_\theta(q(\hat{\theta} + \varepsilon), \hat{\theta} + \varepsilon) > 0$$

and thus (by (2.5))

$$\pi_\theta(\hat{\theta} + \varepsilon) < 0$$

This is only feasible if $\pi(\hat{\theta}) > 0$ and thus $\eta(\hat{\theta}) = 0$. With $\mu(\theta) = 0$ the first order condition (2.10) becomes

$$S - c_q(q(\theta), \theta) + \frac{\lambda(\theta)}{f(\theta)} c_{q\theta}(q(\theta), \theta) = 0 \tag{2.30}$$

Using the implicit function theorem we find

$$q_\theta = \frac{c_{q\theta}(-1 + (\lambda/f)') + c_{q\theta\theta}\lambda/f}{c_{qq} - c_{qq\theta}\lambda/f} \tag{2.31}$$

As derived in the proof of lemma 2.1, $k_\theta = \frac{c_{\theta\theta}}{-c_{q\theta}}$. Comparing q_θ and k_θ in the point $\hat{\theta}$ we can simplify the expression in (2.31) by noting that $c_{q\theta\theta} = c_{qq\theta} = 0$ for $\theta = \hat{\theta}$ by assumption 2.2. Using this we can write $q_\theta(\hat{\theta}) < k_\theta(\hat{\theta})$ as

$$c_{qq}c_{\theta\theta} - c_{q\theta}^2 > c_{q\theta}^2(-\lambda/f)' \tag{2.32}$$

which leads to a contradiction because the left hand side is negative by assumption 2.3 and the right hand side is positive by assumption. Hence, it must be the case that $q_\theta(\hat{\theta}) \geq k_\theta(\hat{\theta})$ at such a point $\hat{\theta}$.

Part 2. Suppose not, that is there exists $\theta'' > \theta'$ such that $c_\theta(q(\theta''), \theta'') > 0$, i.e. such that $q(\theta'') < k(\theta'')$. This implies that there exists $\hat{\theta} \in [\theta', \theta'')$ such that $q(\hat{\theta}) = k(\hat{\theta})$ and $q_\theta(\hat{\theta}) < k_\theta(\hat{\theta})$. Part 1 of this lemma shows that this is not possible.

Part 3. The proof is again by contradiction. Suppose, profits were positive on some interval $(\hat{\theta}_1, \hat{\theta}_2)$ with $\theta_1 < \hat{\theta}_1 < \hat{\theta}_2 < \theta_2$.[22] Quality $q(\theta)$ for $\theta \in (\hat{\theta}_1, \hat{\theta}_2)$ will be determined by (2.30) with $\lambda(\theta) = 1 - F(\theta) - \int_{\hat{\theta}_2}^{\bar{\theta}} \eta(\theta)\, d\theta$. Clearly, there has to be a type $\hat{\theta} \in (\hat{\theta}_1, \hat{\theta}_2)$ at which $\pi(\theta)$ attains a local maximum. Since profits are increasing for $\hat{\theta} - \varepsilon$ and decreasing

[22] By continuity of $\pi(\theta)$, it cannot be the case that $\pi(\theta) > 0$ only in a point.

for $\hat{\theta}+\varepsilon$, (2.5) implies that $q(\hat{\theta}-\varepsilon) > k(\hat{\theta}-\varepsilon)$ and $q(\hat{\theta}+\varepsilon) < k(\hat{\theta}+\varepsilon)$. Hence, $q(\hat{\theta}) = k(\hat{\theta})$ and

$$q_\theta(\hat{\theta}) < k_\theta(\hat{\theta})$$

which is impossible by part 1 of this lemma. This is the required contradiction. *Q.E.D.*

Proof of lemma 2.5 In order to proof this, we need the following result.

Lemma 2.7. *At all qualities greater or equal to his first best quality $q^{fb}(\theta)$, the costs of type θ are lower than the costs of all $\tilde{\theta} < \theta$, i.e. $c(q,\theta) < c(q,\tilde{\theta})\ \ \forall q \geq q^{fb}(\theta)$ and $\tilde{\theta} < \theta$.*

Proof of lemma 2.7 At $q^{fb}(\theta)$ the claim follows from the strictly increasing first best net value assumption (WM): If a lower θ had the same or lower costs at $q^{fb}(\theta)$, he could produce at least the same net value by producing at $q^{fb}(\theta)$. Given that θ has lower costs for $q^{fb}(\theta)$, it is sufficient to show that the incremental costs of producing higher q, i.e. $c(q,\cdot) - c(q^{fb}(\theta),\cdot)$, is lower for θ than for any $\tilde{\theta} < \theta$. Since,

$$c(q,\tilde{\theta}) - c(q^{fb}(\theta),\tilde{\theta}) = \int_{q^{fb}(\theta)}^{q} c_q(x,\tilde{\theta})dx \tag{2.33}$$

is strictly decreasing in $\tilde{\theta}$ because of $c_{q\theta} < 0$, the claim follows. Note that an implication of this claim is that

$$c_\theta(q,\theta) < 0 \text{ for all } q \geq q^{fb}(\theta) \tag{2.34}$$

as otherwise a marginally lower type would have lower costs. *Q.E.D.*

The proof of lemma 2.5 is by contradiction. Suppose there exists θ' such that $\lambda(\theta') < 0$.[23] Since the transversality condition implies $\lambda(\bar{\theta}) = 0$,[24] it follows from the continuity of $\lambda(\theta)$ that there must be an interval of types in between θ' and $\bar{\theta}$ where $\lambda_\theta(\theta) = -f(\theta) + \eta(\theta) > 0$. This can only happen if $\eta(\theta) > 0$ or equivalently $\pi(\theta) = 0$ on this interval. On such a zero profit interval $c_\theta(q,\theta) = 0$ as $\pi_\theta(\theta)$ would not be zero otherwise.[25] Hence, each type produces a quality such that he is the cost minimizing type for this quality. Furthermore, each of these types has $\pi(\theta) = 0$. These two facts imply that incentive compatibility cannot be a problem on the zero profit interval. In other words, we can ignore constraint (SOC) on this interval, i.e. $\mu(\theta) = 0$ on this interval.

[23] As λ is continuous, it is without loss of generality to assume $\theta' > \underline{\theta}$.

[24] Because the highest type has strictly positive profits (see below).

[25] The alternative would be $x(\theta) = 0$. But this is obviously not possible on an interval of types by assumption 2.1.

Denote the lowest type with zero profits as $\theta_1 = \inf\{\theta|\pi(\theta) = 0, \theta \geq \theta_1\}$. Note that from what was said above $\lambda(\theta_1) < 0$ and $\mu(\theta_1) = 0$. Furthermore, $c_\theta(q(\theta_1 - \varepsilon), \theta_1 - \varepsilon) \geq 0$ for $\varepsilon > 0$ small enough.[26] Equation (2.34) then implies $q(\theta_1 - \varepsilon) < q^{fb}(\theta_1 - \varepsilon)$. However, this contradicts the first order condition with respect to q:

$$f(\theta)(S_q(q(\theta)) - c_q(q(\theta), \theta)) + \lambda(\theta)c_{q\theta}(q(\theta), \theta) + \mu(\theta)c_{q\theta\theta}(q(\theta), \theta) = -\mu_\theta(\theta)c_{q\theta}(q(\theta), \theta) \tag{2.35}$$

Since $\mu(\theta_1) = 0$, we can ignore the $\mu(\theta_1 - \varepsilon)$ term for $\varepsilon > 0$ small enough. Further, $\mu_\theta(\theta_1 - \varepsilon) \leq 0$ since $\mu(\theta) \geq 0$. But then $c_{\theta q} < 0$ and $\lambda(\theta_1 - \varepsilon) < 0$ imply $S_q(q(\theta_1 - \varepsilon)) - c_q(q(\theta_1 - \varepsilon), \theta_1 - \varepsilon) < 0$ which contradicts $q(\theta_1 - \varepsilon) < q^{fb}(\theta_1 - \varepsilon)$. Hence, there is a contradiction and $\lambda(\theta) \geq 0$ has to hold.

To prove the other part of the lemma, suppose (again by contradiction) that $\pi(\underline{\theta}) > 0$. Consequently, the dynamic optimization problem will include the transversality condition $\lambda(\underline{\theta}) = 0$. Given that $\lambda_\theta(\theta) = -f(\theta) + \eta(\theta)$, this implies that $\lambda(\theta) < 0$ for some interval of θ starting at $\underline{\theta}$.[27] As we just proved, it is not possible to have $\lambda(\theta) < 0$. This is the required contradiction and we conclude that $\pi(\underline{\theta}) = 0$. *Q.E.D.*

Proof of proposition 2.1

We will analyze the problem without the incentive constraint (SOC) first, i.e. $\mu(\theta) = 0$, and show afterwards in lemma 2.9 that it is satisfied. The first order condition (2.10) becomes

$$S - c_q(q(\theta), \theta) + \frac{\lambda(\theta)}{f(\theta)}c_{q\theta}(q(\theta), \theta) = 0 \tag{2.36}$$

Since $\bar{\theta}$ is the best type (in a first best sense), we expect his profits to be positive and therefore $\eta(\bar{\theta}) = 0$ and also the transversality condition $\lambda(\bar{\theta}) = 0$ holds (indeed below we verify that $\pi(\bar{\theta}) > 0$). Therefore (2.9) implies $\lambda(\theta) = 1 - F(\theta)$ for some high types for which the profit constraint does not bind. Note that for this case, equation (2.36) can be written as (2.13).

Now we have two cases. With the solution $q^h(\theta)$ given by (2.13) it is the case that either

1. $c_\theta(q^h(\underline{\theta}), \underline{\theta}) < 0$ or

[26] This follows form the definition of θ_1: Since $\pi(\theta_1) = 0$ and $\pi(\theta_1 - \varepsilon) \geq 0$, profits have to be decreasing at $\theta_1 - \varepsilon$ for ε small enough.

[27] To be precise, this follows from the continuity of $\pi(\theta)$: As $\pi(\underline{\theta}) > 0$, profits have to be positive for some interval of low θ and consequently $\eta(\theta) = 0$ for those θ. This implies $\lambda(\theta) < 0$.

2. $c_\theta(q^h(\underline{\theta}), \underline{\theta}) \geq 0$

The first case implies that $q(\underline{\theta}) > k(\underline{\theta})$. Hence, for the first case, $\pi(\underline{\theta}) = 0$ and $\pi_\theta(\theta) > 0$ for (at least) θ close to $\underline{\theta}$ (see equation (2.5)). It follows from part 2 of lemma 2.4 that $c_\theta(q(\theta), \theta) \leq 0$ for all θ. Thus $\pi_\theta \geq 0$ for each $\theta > \underline{\theta}$ and the profit constraint $\pi(\theta) \geq 0$ does not bind for $\theta > \underline{\theta}$. Therefore the solution in equation (2.13) is the overall solution.

Finally, consider the virtual surplus in equation (2.11) with $\lambda(\theta) = 1 - F(\theta)$. Using an envelope argument, it is routine to derive that

$$\frac{dVV(\theta)}{d\theta} = -c_\theta(1 - \left(\frac{1-F}{f}\right)') + \frac{1-F}{f}c_{\theta\theta} > 0 \tag{2.37}$$

Since the project is allocated to the firm with the highest VV, it is allocated to the firm with the highest θ. Thus $x_\theta(\theta) > 0$.

Now consider the second case in proposition 2.1 with $c_\theta(q^h(\underline{\theta}), \underline{\theta}) \geq 0$.

Lemma 2.8. *If $c_\theta(q^h(\underline{\theta}), \underline{\theta}) \geq 0$, then $q(\underline{\theta}) = k(\underline{\theta})$.*

Proof of lemma 2.8 If $c_\theta(q^h(\underline{\theta}), \underline{\theta}) = 0$, the lemma is true.

If $c_\theta(q(\underline{\theta}), \underline{\theta}) > 0$ then $\pi_\theta(\underline{\theta}) < 0$. Then $\pi(\underline{\theta}) = 0$ (lemma 2.5) implies that this violates the constraint that profits should be non-negative. In this case the solution cannot be given by equation (2.13) as the profit constraint is binding. Hence, the solution $q(\theta)$ is given by equation (2.30) where (see equation (2.9)) $\lambda(\theta)$ is given by equation (2.28). This solution cannot feature $c_\theta(q(\underline{\theta}), \underline{\theta}) > 0$ as this would lead to a violation of $\pi(\theta) \geq 0$ for θ close to $\underline{\theta}$.

The following argument shows that $c_\theta(q(\underline{\theta}), \underline{\theta}) < 0$ is not possible either. In this case $\pi_\theta(\underline{\theta}) > 0$. Then either (i) there exists $\theta' > \underline{\theta}$ such $\pi(\theta') = 0$ or (ii) $q(\theta) = q^h(\theta)$. Case (i) leads to a contradiction because of lemma 2.4. If (i) does not happen, then $\eta(\theta) = 0$ for all $\theta > \underline{\theta}$, which implies case (ii). However, case (ii) with $c_\theta(q(\underline{\theta}), \underline{\theta}) < 0$ contradicts the assumption in the lemma that $c_\theta(q^h(\underline{\theta}), \underline{\theta}) \geq 0$.

Thus we have $c_\theta(q(\underline{\theta}), \underline{\theta}) = 0$ or equivalently $q(\underline{\theta}) = k(\underline{\theta})$. *Q.E.D.*

Because of lemma 2.8, there is a largest $\theta_b \geq \underline{\theta}$ such that $q(\theta) = k(\theta)$ for all $\theta \in [\underline{\theta}, \theta_b]$. This θ_b is uniquely defined.

Since $\pi(\underline{\theta}) = 0$ and $\pi_\theta = 0$ for all $\theta \in [\underline{\theta}, \theta_b]$, we have $\pi(\theta) = 0$ for all $\theta \in [\underline{\theta}, \theta_b]$.

Uniqueness of θ_b as defined in (2.15) follows from the fact that the expression in equation (2.15) is strictly increasing in θ_b. Differentiating the expression with respect to θ_b and using assumption 2.2 we find:

$$\frac{1}{-c_{q\theta}}(c_{q\theta}^2 - c_{qq}c_{\theta\theta} - c_{q\theta}^2\left(\frac{1-F}{f}\right)') > 0$$

We can only leave the interval $[\underline{\theta}, \theta_b]$ if $\pi_\theta(\theta_b + \varepsilon) > 0$ for $\varepsilon > 0$ small enough. Then $\pi(\theta) > 0$ for all $\theta > \theta_b$. If not, there would be $\theta' > \theta_b$ such that $\pi(\theta') = 0$ which contradicts lemma 2.4. Hence, $q(\theta) = q^h(\theta)$ for all $\theta > \theta_b$ and equation (2.15) makes sure that $q(\theta)$ is continuous.

As in the previous case, we have $q_\theta(\theta) > 0$ for $\theta > \theta_b$. For $\theta \in [\underline{\theta}, \theta_b]$ we have $q(\theta) = k(\theta)$ which is (strictly) increasing in θ if $c_{\theta\theta} > 0$. If $c_{\theta\theta} = 0$, quality is constant over the range $\theta \in [\underline{\theta}, \theta_b]$.

Finally, we show that $x_\theta(\theta) \geq 0$:

$$\frac{dVV(\theta)}{d\theta} = -c_\theta(1 - \left(\frac{\lambda(\theta)}{f}\right)') + \frac{\lambda(\theta)}{f}c_{\theta\theta} \geq 0 \tag{2.38}$$

where the inequality is strict for $\theta > \theta_b$ and for $\theta \in [\underline{\theta}, \theta_b]$ if $c_{\theta\theta} > 0$.

Finally, lemma 2.9 establishes global incentive compatibility. *Q.E.D.*

Lemma 2.9. *The relaxed solution in proposition 2.1 is globally incentive compatible.*

Proof of lemma 2.9 The monotonicity of $x(\theta)$ and $q(\theta)$ together with $c_\theta \leq 0$ and $c_{q\theta} < 0$ imply that the local incentive compatibility constraint (SOC) is satisfied.

For global incentive compatibility we first show that no θ can profitably misrepresent as $\hat{\theta} > \theta$. This is true if

$$\pi(\theta) - \pi(\hat{\theta}) - x(\hat{\theta})[c(q(\hat{\theta}), \hat{\theta}) - c(q(\hat{\theta}), \theta)] \geq 0$$

Using (2.5), this can be rewritten as

$$\int_\theta^{\hat{\theta}} x(t)c_\theta(q(t), t) - x(\hat{\theta})c_\theta(q(\hat{\theta}), t)\, dt \geq 0$$

This last inequality can be rewritten as

$$\int_\theta^{\hat{\theta}} \int_t^{\hat{\theta}} x_\theta(s)c_\theta(q(s), t) + x(s)c_{q\theta}(q(s), t)q_\theta(s)\, ds\, dt \leq 0 \tag{2.39}$$

The second term of the integrand is negative by the monotonicity of $q(\theta)$ in proposition 2.1. Note that we saw in the proof of proposition 2.1 that $c_\theta(q(\theta), \theta) \leq 0$ for all types.

Since $t \leq s$ and $c_{\theta\theta} \geq 0$, clearly $c_t(q(s), t) \leq 0$ in the first term of the integrand. As $x_\theta \geq 0$ in proposition 2.1, inequality (2.39) has to hold.

To show that no θ gains by misrepresenting as $\hat{\theta} < \theta$ we use the following notation introduced in equation (2.3).

$$\pi(\hat{\theta}, \theta) = t(\hat{\theta}) - x(\hat{\theta})c(q(\hat{\theta}), \theta)$$

The idea is to define the following cost function

$$\tilde{c}(a, \theta) = \min\{c(q(a), a), c(q(a), \theta)\} \tag{2.40}$$

where $q(a)$ is the optimal quality schedule derived above. Next define

$$\tilde{\pi}(a, \theta) = t(a) - x(a)\tilde{c}(a, \theta) \tag{2.41}$$

The following inequalities show that the solution derived above satisfies IC globally as well:

$$\begin{aligned}
& \pi(\hat{\theta}, \theta) - \pi(\theta, \theta) \\
\leq\ & \tilde{\pi}(\hat{\theta}, \theta) - \tilde{\pi}(\theta, \theta) \\
=\ & \int_\theta^{\hat{\theta}} \frac{\partial \tilde{\pi}(a, \theta)}{\partial a} da \\
=\ & \int_{\hat{\theta}}^{\theta} \left(\left. \frac{\partial \pi(a, \theta)}{\partial a} \right|_{\theta=a} - \frac{\partial \tilde{\pi}(a, \theta)}{\partial a} \right) da \qquad (2.42) \\
=\ & \int_{\hat{\theta}}^{\theta} x_\theta(a)(\tilde{c}(q(a), \theta) - c(q(a), a)) + x(a)(\tilde{c}_a(q(a), \theta) - c_q(q(a), a)q_\theta(a))da \qquad (2.43) \\
\leq\ & 0
\end{aligned}$$

where the first inequality follows from the definition of $\tilde{c}(\cdot)$ and the observation that $\tilde{\pi}(\theta, \theta) = \pi(\theta, \theta)$. Equation (2.42) follows because $\left. \frac{\partial \pi(a,\theta)}{\partial a} \right|_{\theta=a} = 0$ by the first order condition of truthful revelation. Equation (2.43) follows from the definitions of the derivatives of $\pi(a, \theta)$ and $\tilde{\pi}(a, \theta)$ w.r.t. a. The final inequality follows from the properties $x_a(a), q_a(a) \geq 0$ and the following three observations. First, by definition of $\tilde{c}(\cdot)$ we have

$$\tilde{c}(q(a), \theta) - c(q(a), a) \leq 0$$

Second, for values of a where $\tilde{c}(a, \theta) = c(q(a), \theta)$ we have

$$\tilde{c}_a(q(a), \theta) - c_q(q(a), a)q_a(a) = (c_q(q(a), \theta) - c_q(q(a), a))q_a(a) \leq 0$$

because $c_{q\theta} \leq 0$. Finally for values where $\tilde{c}(a,\theta) = c(q(a),a)$ we have

$$\tilde{c}_a(q(a),\theta) - c_q(q(a),a)q_a(a) = \left.\frac{\partial c(q(a),\theta)}{\partial \theta}\right|_{\theta=a} \leq 0$$

because in our solution $c_\theta(q(\theta),\theta) \leq 0$ for all θ. *Q.E.D.*

Proof of lemma 2.6 We show that a menu featuring $\pi(\underline{\theta}) = 0$ is not optimal. In the WNM case, first best welfare is decreasing in type around $\underline{\theta}$. A standard envelope argument shows that this implies $c_\theta(q^{fb}(\underline{\theta}),\underline{\theta}) > 0$. Now suppose, $\pi(\underline{\theta}) = 0$. Then $\pi_\theta(\underline{\theta}) \geq 0$ which implies $c_\theta(q(\underline{\theta}),\underline{\theta}) \leq 0$ by (2.5). Therefore, $q(\underline{\theta}) > q^{fb}(\underline{\theta})$. But then a simple change in the menu would be beneficial and therefore the menu cannot be optimal: Change $q(\underline{\theta})$ to $q^{fb}(\underline{\theta})$ and adjust transfers such that $\pi(\underline{\theta})$ stays zero. As $\underline{\theta}$ has again zero profits his incentive compatibility does not change. Reducing quality will make $\underline{\theta}$'s menu point even less attractive for other types.[28] By the definition of $q^{fb}(\cdot)$, this change is beneficial. *Q.E.D.*

Proof of proposition 2.2 Again the global IC constraint will be neglected first and checked ex post.

From lemma 2.6 we know that $\pi(\bar{\theta}), \pi(\underline{\theta}) > 0$ and therefore the transversality conditions $\lambda(\bar{\theta}) = \lambda(\underline{\theta}) = 0$ have to hold. Furthermore, the positive profit constraint is non binding and therefore $\eta(\bar{\theta}) = \eta(\underline{\theta}) = 0$. By (2.9) and the continuity of π, we have $\lambda(\theta) = 1 - F(\theta) > 0$ close to $\bar{\theta}$ and $\lambda(\theta) = -F(\theta) < 0$ close to $\underline{\theta}$. For these two expressions of $\lambda(\cdot)$, the monotone hazard rate assumption implies that the quality schedule determined in (2.30) is increasing in type, i.e. $q_\theta(\theta) > 0$.

Next we proof the existence of θ_1 and θ_2. By continuity of $\lambda(\theta)$,[29] there exists an interval $[\tilde{\theta}_1, \tilde{\theta}_2]$ such that $\lambda_\theta(\theta) = -f(\theta) + \eta(\theta) > 0$ and thus $\pi(\theta) = 0$. Consequently, $\pi_\theta(\theta) = 0$ for all $\theta \in [\tilde{\theta}_1, \tilde{\theta}_2]$. Let θ_1 (θ_2) denote the lowest (highest) $\tilde{\theta}_1$ ($\tilde{\theta}_2$) such that this is true for all $\theta \in [\theta_1, \theta_2]$. By continuity of $q(\theta)$ it follows that $q^l(\theta_1) = k(\theta_1)$ and $q^h(\theta_2) = k(\theta_2)$.

As shown in the proof of proposition 2.1 the expression in equation (2.15) is strictly increasing in θ_b. This implies the uniqueness of $\theta_2 = \theta_b$. With a similar argument one shows that

$$S - c_q(k(\theta),\theta) - \frac{F(\theta)}{f(\theta)} c_{q\theta}(k(\theta),\theta)$$

[28] This argument can be made formal using the same dc expression as in the proof of lemma 2.7.

[29] In particular, to connect $\lambda(\theta) < 0$ for small θ with $\lambda(\theta) > 0$ for high θ, we need $\lambda_\theta > 0$ over some range.

is increasing in θ. This implies the uniqueness of θ_1 which solves

$$S - c_q(k(\theta_1), \theta_1) - \frac{F(\theta_1)}{f(\theta_1)} c_{q\theta}(k(\theta_1), \theta_1) = 0$$

Since

$$S - c_q(k(\theta), \theta) - \frac{F(\theta)}{f(\theta)} c_{q\theta}(k(\theta), \theta) > S - c_q(k(\theta), \theta) + \frac{(1 - F(\theta))}{f(\theta)} c_{q\theta}(k(\theta), \theta)$$

for all θ it follows that indeed $\theta_1 < \theta_2$.

By the uniqueness of θ_1 and θ_2, c_θ is positive for $\theta < \theta_1$ and negative for $\theta > \theta_2$. Together with (2.5) this implies the sign of π_θ as stated in the proposition.

In (θ_1, θ_2), there has to be a type with $\lambda(\theta) = 0$. From (2.30), this type produces his first best quality and as he is in the zero profit interval $q^{fb}(\theta) = k(\theta)$. The only type satisfying this conditions is the type with the lowest first best welfare θ_w. Note that all $\theta < (>)\theta_w$ have $c_\theta(q(\theta), \theta) \geq (\leq)0$ and also $\lambda(\theta) < (>)0$. Differentiating the virtual valuation with respect to θ tells us the sign of x_θ:

$$\frac{d\,VV}{d\theta} = c_\theta(q(\theta), \theta) \left(-1 + \frac{\partial\, \lambda(\theta)/f(\theta)}{\partial \theta} \right) + \frac{\lambda(\theta)}{f(\theta)} c_{\theta\theta}(q(\theta), \theta) \tag{2.44}$$

From the paragraph above and the monotone hazard rate assumption, the virtual valuation, and therefore $x(\theta)$, has to be decreasing on $[\underline{\theta}, \theta_1]$ and increasing on $[\theta_2, \bar{\theta}]$. On (θ_1, θ_2), c_θ is zero and as λ flips sign at θ_w the proposition follows.

It was already mentioned that $q(\theta)$ is increasing for types with positive profits. Since $k(\theta)$ is non-decreasing, $q(\theta)$ is non-decreasing for all θ.

Finally, lemma 2.10 establishes global incentive compatibility. *Q.E.D.*

Lemma 2.10. *The relaxed solution in proposition 2.2 is globally incentive compatible.*

Proof of lemma 2.10 All $\theta \in [\theta_1, \theta_2]$ produce at $k(\theta)$ which is the quality level at which a type has lower cost than any other type. Since these types also have zero profits, no other type can profitably misrepresent as $\theta \in [\theta_1, \theta_2]$. For $\theta \geq \theta_w$ the menu is equivalent to the one described in proposition 2.1. Therefore, lemma 2.9 implies non-local IC on this part of the menu. The same proof as for lemma 2.9 with reversed signs implies that the menu for $\theta < \theta_w$ is non-locally IC.

What remains to be shown is that no type $\theta < \theta_w$ can profitably misrepresent as $\theta' > \theta_w$ (and the other way round). Take such a θ and observe that θ_2 has lower costs at $q(\theta')$:

$$c(q(\theta'), \theta_2) - c(q(\theta'), \theta) = \int_\theta^{\theta_2} c_\theta(q(\theta'), t)\, dt < 0 \tag{2.45}$$

The inequality follows from the fact that $k(\theta), k(\theta_2) < q(\theta')$ and $c_{q\theta} < 0$. Therefore, the integrand is negative over the whole range. Incentive compatibility for θ requires

$$\begin{aligned}\pi(\theta) &\geq \pi(\theta') + x(\theta')[c(q(\theta'),\theta') - c(q(\theta'),\theta)] \\ &= \underbrace{\pi(\theta') + x(\theta')[c(q(\theta'),\theta') - c(q(\theta'),\theta_2)]}_{\leq 0} + x(\theta')[c(q(\theta'),\theta_2) - c(q(\theta'),\theta)].\end{aligned}$$

The first term in the last expression is negative because incentive compatibility between θ_2 and θ' is satisfied (see lemma 2.9 and recall that $\pi(\theta_2) = 0$). The second term is also negative because of equation (2.45). As $\pi(\theta) \geq 0$, the inequality above and therefore incentive compatibility holds.

The proof for $\theta > \theta_w$ and $\theta' < \theta_w$ works in the same way with θ_1 in place of θ_2. *Q.E.D.*

Proof of proposition 2.3 Consider $\theta' = \underline{\theta}$. Define $\underline{W} = W^{fb}(\underline{\theta}) = W^{sb}(\underline{\theta})$. Since $\underline{\theta}$ produces his first best quality and first best welfare is decreasing at $\underline{\theta}$, there are types $\theta > \underline{\theta}$ with lower welfare than $\underline{W}$. By the definition of the (WNM)-case, $W^{fb}(\bar{\theta}) > \underline{W}$.

Taking these two points together and applying the intermediate value theorem yields the existence of a type θ'' such that $W^{sb}(\theta'') = \underline{W}$ and $W^{sb}_{\theta}(\theta'') > 0$.

$$\frac{dW^{sb}(\theta)}{d\theta} = (S - c_q(q(\theta),\theta))q_\theta(\theta) - c_\theta(q(\theta),\theta) = -\frac{\lambda(\theta)}{f(\theta)}c_{q\theta}(q(\theta),\theta)q_\theta(\theta) - c_\theta(q(\theta),\theta)$$

where the first order condition for $q(\cdot)$ is used for the second equality. From proposition 2.2 and its proof we know that λ and c_θ both change sign at θ_w and therefore $sign(\lambda(\theta)) = -sign(c_\theta(q(\theta),\theta))$. Consequently, $W^{sb}_{\theta}(\theta'') > 0$ implies $\lambda(\theta'') > 0$ and $c_\theta(q(\theta''),\theta'') < 0$.

The virtual valuation can be written as

$$VV(\theta) = W^{sb}(\theta) + \frac{\lambda(\theta)}{f(\theta)}c_\theta(q(\theta),\theta)$$

and thus $VV(\theta) \leq W^{sb}(\theta)$ since λ and c_θ have opposite signs and the inequality is strict if $\lambda(\theta), c_\theta(q(\theta),\theta) \neq 0$.

If $c_\theta(q(\theta''),\theta'') < 0$ it then follows that $VV(\underline{\theta}) > VV(\theta'')$. By continuity, there exist types θ that yield strictly higher welfare than $\underline{\theta}$ but still lose from $\underline{\theta}$ in the procurement.

Now consider the case where $\theta'' \in (\theta_1, \theta_2)$ such that $c_\theta(q(\theta''),\theta'') = 0$. In this case there are types close to $\underline{\theta}$ that lose from types close to θ'' although the former yield higher (second best) welfare W^{sb}. *Q.E.D.*

Proof of proposition 2.4: If the scoring rule implements the optimal mechanism it has to hold that $bid(\theta') = bid(\theta'')$ whenever $VV(\theta') = VV(\theta'')$ under the optimal mechanism.

Now take θ_1 and θ_2 as defined in proposition 2.2. Because $bid_\theta(\theta) = -c_\theta(q(\theta), \theta)$ and all $\theta \in (\theta_1, \theta_2)$ have $q(\theta) = k(\theta)$, it follows that $bid(\theta_1) = bid(\theta_2)$. As virtual valuation and bids are continuous in type, this implies that $VV(\theta_1) = VV(\theta_2)$ has to hold if the scoring rule implements the optimal mechanism: Otherwise, types slightly below θ_1 and slightly above θ_2 have the same bid but different virtual valuations. Since $q(\theta_i) = k(\theta_i)$, the virtual valuation for θ_i is $Sk(\theta_i) - c(k(\theta_i), \theta_i)$ for $i = 1, 2$. Consequently, the following equation has to hold if the scoring rule implements the optimal mechanism:

$$\int_{\theta_1}^{\theta_2} \frac{d\{Sk(\theta) - c(k(\theta), \theta)\}}{d\theta}\, d\theta = 0$$

This can be rewritten as

$$\int_{\theta_1}^{\theta_2} \frac{(S - c_q(k(\theta), \theta))c_{\theta\theta}(k(\theta), \theta)}{-c_{q\theta}(k(\theta), \theta)}\, d\theta = 0.$$

Note that this equation uniquely pins down θ_2 for a given θ_1.[30] Furthermore, it does so independent of the distribution of types. However,θ_2 is defined by the equation $S - c_q(k(\theta), \theta) + \frac{1-F(\theta)}{f(\theta)c_{q\theta}k(\theta),\theta)}$ which depends on $f(\theta_2)$. Hence, slightly perturbing f around θ_2 changes θ_2 but not the equation above. Consequently, a scoring rule auction cannot implement the optimal mechanism in a generic sense. *Q.E.D.*

Proof of proposition 2.5 As shown in the proof of lemma 2.9, incentive compatibility between θ and $\hat{\theta}$ boils down to the inequality

$$\int_{\theta}^{\hat{\theta}} \int_{t}^{\hat{\theta}} x_\theta(s)c_\theta(q(s), t) + x(s)c_{\theta q}(q(s), t)q_\theta(s)\, ds\, dt \leq 0.$$

Now note that $c_{\theta\theta} = 0$ implies

$$x_\theta(s)c_\theta(q(s), t) + x(s)c_{\theta q}(q(s), t)q_\theta(s) = x_\theta(s)c_\theta(q(s), s) + x(s)c_{\theta q}(q(s), s)q_\theta(s).$$

But then global incentive compatibility has to be satisfied as $x_s(s)c_\theta(q(s), s) + x(s)c_{\theta q}(q(s), s)$ $q_s(s) \leq 0$ by the local second order condition. *Q.E.D.*

[30] The reason is that the integrand is negative around θ_1, positive around θ_2 and changes sign only at one type which is between θ_1 and θ_2. This follows from lemma 2.1.

Chapter 3

Health insurance without single crossing: Why healthy people have relatively high coverage[31]

3.1. Introduction

A well documented problem in health insurance markets with voluntary insurance (like the US) is that people either have no insurance at all or are underinsured.[32] Standard insurance models (inspired by the seminal work of Rothschild and Stiglitz (1976) (RS) and Stiglitz (1977)) predict that *healthy* people have less than perfect insurance or –in the extreme– no insurance at all. However, both popular accounts like Cohn (2007) and academic work like Schoen et al. (2008) show that people with low health status are overrepresented in the group of uninsured and underinsured.[33] We develop a model to explain why sick people end up with little or no insurance. We do this by adding two well documented empirical observations (discussed below) to the RS model: (i) richer people tend to be healthier and (ii) health is a normal good. Technically speaking, introducing the latter two effects can lead to a violation of single crossing in the model.

Another indication that the standard RS framework (with single crossing) does not

[31] This chapter is based on Boone and Schottmüller (2011a).

[32] In empirical studies, underinsurance is defined using indicators of financial risk. To illustrate, one definition of underinsurance used by Schoen et al. (2008) is "out-of-pocket medical expenses for care amounted to 10 percent of income or more". In our theoretical model, underinsurance refers to less than socially optimal/efficient insurance.

[33] In the words of Schoen et al. (2008, pp. w303): "underinsurance rates were higher among adults with health problems than among healthier adults".

capture reality well is the following. The empirical literature that is based on RS does not unambiguously show that asymmetric information plays a role in health insurance markets. One would expect people to have private information about their health risks– think for example of preconditions, medical history of parents and other family members or life style. However, some papers, like for example Cardon and Hendel (2001) or Cutler et al. (2008), do not find evidence of asymmetric information while others do, e.g. Bajari et al. (2005) or Munkin and Trivedi (2010). The test for asymmetric information employed in these papers is the so called "positive correlation test," i.e. testing whether riskier types buy insurance contracts with higher coverage.[34]

We show that an insurance model with a violation of single crossing is capable of explaining why healthy people have better insurance (in equilibrium) than people with a low health status. In particular, the positive correlation property no longer holds if single crossing is violated. Consequently, testing for this positive correlation can no longer be viewed as a test for asymmetric information. As mentioned, we use two well documented stylized facts to motivate this violation of single crossing in the market for health insurance.

Single crossing means that people with higher health risks have a higher willingness to pay for marginally increasing coverage, e.g. reducing copayments. If this property holds for all possible coverage levels, a given indifference curve of a high risk type can cross a given indifference curve of a low risk type at most once. A rough intuition for why the stylized facts above can lead to a violation of single crossing is given by the following: At full coverage (indemnity insurance that pays for all medical costs), high risk (low health) types will tend to spend more on treatments than low risk types. Hence, a small reduction in coverage, leads to a bigger loss in utility for high risk types. Now consider health insurance with low coverage where the insured faces substantial copayments. Because health is a normal good, it is possible that the rich-healthy type spends more on treatment than the low income, low health type. Put differently, a rich-healthy type might utilize the insurance more conditional on falling ill. In that case, a small change in coverage can have a bigger effect on the utility of the healthy type than of the low health agent. The healthy type will therefore have a higher willingness to

[34] "Risk" is in structural estimation papers–broadly speaking–interpreted as a parameter on which the distribution of health shocks depends.

pay for a marginal increase in coverage than the low health type. This violates single crossing.

We show that in insurance models without single crossing higher health risks are not necessarily associated with more coverage while this prediction is inevitable with single crossing.

The starting point for our paper is the positive correlation property which is established in various forms in the theoretical literature. The most general treatment is Chiappori et al. (2006). However, their main focus is a positive correlation between coverage and insurance payout while we are more interested in the correlation between patient risk and coverage.[35] On a technical level, their assumption (NIP) does not hold in our setting when firms have market power.

The literature on violations of single crossing is relatively scarce. There are three papers analyzing perfectly competitive insurance markets with 2×2 types: People differ in two dimensions and both dimensions can either take a high or a low value. In Smart (2000) and Wambach (2000) the two dimensions are risk and risk aversion. Netzer and Scheuer (2010) model an additional labor supply decision and the two dimensions are productivity and risk. All papers have a pooling result, i.e. if single crossing does not hold two of the four types can be pooled. Only in Netzer and Scheuer (2010) there can be equilibria where some low risk types have more coverage than some high risk types. Nevertheless, also in their paper the type having the highest coverage in equilibrium is a high risk type. Furthermore, in their model the wealthiest types have the lowest coverage. This contrasts with the empirical observation in the health sector mentioned above.[36] In Smart (2000) and Wambach (2000) the high risk/high risk aversion type receives full coverage and the (low, low) type gets partial coverage. The $(high, low)$ and $(low, high)$ type can be pooled on an intermediate coverage level. Although two types with different risks are pooled, the positive correlation property still holds (weakly) in those models. The pooling itself is a result of the fact that some high risk types are less risk averse than some low risk types. Given that high risk types are likely to be poor

[35] As we argue below, when single crossing is violated, higher risk types do not necessarily have higher insurance payout.

[36] On a technical level, agents in Netzer and Scheuer (2010) make an endogenous decision (labor supply) before the risk realizes. We focus on the decision how much to spend on treatment which is made after the risk realizes.

in the health insurance context, even this pooling result does not arise naturally in the health insurance sector.

Another contribution of our paper is to deviate from the perfect competition assumption. We show that under imperfect competition types cannot only be pooled but high risk types might get less coverage in equilibrium than low risk types. Under perfect competition this is not the case in our model: There would always be a profitable pooling contract in such a situation. If firms have market power, they do not offer this pooling contract as this will reduce their profits from low risk types.

Jullien et al. (2007) take a different approach to answer the question why high risk types might have lower coverage in general insurance markets. They use a model where types differ in risk aversion and single crossing is satisfied. Hence, types with higher risk aversion will have more coverage in equilibrium. At the same time more risk averse agents might engage more in preventive behavior. If types are still separated in equilibrium and risk aversion differences remain the driving force, high risk aversion types will exhibit less risk (due to prevention) and higher coverage. The differences in risk could together with the differences in risk aversion lead to a violation of single crossing, see Araujo and Moreira (2003) for a model of this.[37] We contribute by explicitly analyzing a framework where single crossing is violated and show how different modes of competition on the supply side affect market outcomes. Similar explanations for "advantageous selection" as in Jullien et al. (2007) can be found in Hemenway (1990) and De Meza and Webb (2001).

Since risk in the health sector is exogenously different for different persons (e.g. due to genetics),[38] we follow RS and take a different starting point than Jullien et al. (2007). We assume risk differences instead of risk aversion differences. The result that high risk people have low coverage is in our paper not the result of low risk aversion. The driving force is the violation of single crossing caused by empirically documented income differences between high risks and low risks (see section 3.3). This is also in line with, for example, Fang et al. (2008) who show for the medigap insurance market that income

[37] This is not analyzed in Jullien et al. (2007) as single crossing holds in their setup with CARA utility function.

[38] Also, with the observed correlation between coverage and income in the health insurance market, the assumption by Jullien et al. (2007) implies that high income people are more risk averse than low income people. It is not clear that this is a reasonable assumption in our health insurance context.

is a source of advantageous selection while risk aversion is not.

In the following section, a general insurance model is introduced and equilibrium results for perfect competition, monopoly and oligopoly settings are derived without assuming single crossing. Section 3.3 explains why single crossing is likely to be violated in the health insurance market. In section 3.4, setup and results are illustrated with a numerical example. Section 3.5 concludes. Proofs are relegated to the appendix.

3.2. Insurance model

This section introduces a general model of (health) insurance that allows us to consider both the case where single crossing (SC) is satisfied and the case where it is not satisfied (NSC). After describing the demand side of the insurance market, we consider three alternatives for the supply side: perfect competition, monopoly and oligopoly.

3.2.1. Demand side model

Following RS, we consider an agent with utility function $u(q, p, \theta)$ where $q \in [0, 1]$ denotes coverage or generosity of her insurance contract,[39] $p \geq 0$ denotes the price of insurance (insurance premium) and $\theta \in \{\theta^l, \theta^h\}$ with $\theta^h > \theta^l > 0$ denotes the type of consumer.[40] Higher θ denotes a higher risk in the sense of higher expected costs (in case $q^h = q^l = 1$; see below). This could, for instance, be the case due to chronic illness or higher risk due to a genetic precondition. We make the following assumptions on the utility function (where subscripts denote partial derivatives).

Assumption 3.1. *The utility function $u(q, p, \theta)$ is continuous and differentiable. It satisfies $u_q > 0, u_p < 0$. We define the indifference curve $p(q, u, \theta)$ as follows:*

$$u(q, p(q, u, \theta), \theta) \equiv u \tag{3.1}$$

[39] Apart from literal coverage –where $1 - q$ denotes the agent's copayments– q could, for example, be interpreted as $1/(1 + \text{deductible})$. Note that in models without moral hazard both parameters are similar in the sense that high risk types dislike co-payments and deductibles more relative to low risk types.

[40] We follow RS in assuming that there are only two types. For an analysis of a violation of single crossing with a continuum of types θ, see Araujo and Moreira (2010) and Schottmüller (2011a).

We assume that these indifference curves $p(q, u, \theta)$ *are differentiable in* q *and* u *with* $p_q = -u_q/u_p > 0, p_u = 1/u_p < 0$.
Further, the crossing at $q = 1$ *satisfies:*

$$p_q(1, u^h, \theta^h) > p_q(1, u^l, \theta^l) \tag{C1}$$

for all $u^l \geq \bar{u}^l = u(0, 0, \theta^l), u^h \geq \bar{u}^h = u(0, 0, \theta^h)$.

In words, utility u is increasing in coverage q and decreasing in the premium p paid for insurance. For given type θ and utility level u, the indifference curve $p(q, u, \theta)$ maps out combinations (q, p) that yield the same utility. Because higher coverage leads to higher utility, p has to increase to keep utility constant. Hence, indifference curves are upward sloping in (q, p) space ($p_q > 0$). Increasing u (for a given coverage level q) requires a lower price. Thus, raising u shifts an indifference curve downwards ($p_u < 0$).

Type $k \in \{h, l\}$ buys insurance if it leads to a higher utility than her outside option $\bar{u}^k$. This outside option is given by the "empty insurance contract": $q = p = 0$.

At $q = 1$ a marginal reduction in q should be compensated by a bigger decrease in p for θ^h compared to θ^l. This reflects the fact that the θ^h type faces higher expected health care expenditures, i.e. he is the high risk type. At $q = 1$, i.e. at full coverage, other factors like willingness to pay for treatment (which could be different for different types) do not play a role. In this sense, this assumption "defines" what higher θ means: at $q = 1$, higher θ types face higher expected costs. With the same idea we assume that expected costs for the insurer of a contract with $q = 1$ is higher for the θ^h than for the θ^l type: $c(1, u^h, \theta^h) > c(1, u^l, \theta^l)$ for all $u^h \geq \bar{u}^h, u^l \geq \bar{u}^l$. Intuitively, u should not matter for health care consumption at full coverage and the high risk type will use the insurance more.

To allow for income effects (for instance, in treatment choice; see below) the cost function depends on u. However, we assume two regularity conditions.

Assumption 3.2. *For each type* $k \in \{h, l\}$ *and* $q \in [0, 1]$ *we assume that*

- $c_u(q, u^k, \theta^k) \geq 0$ *for* $u^k \geq \bar{u}^k$,
- $c(1, u^k, \theta^k) = c(1, \tilde{u}^k, \theta^k)$, *for* $u^k, \tilde{u}^k \geq \bar{u}^k$.

In words, as the income of the agent increases (which *ceteris paribus* leads to higher utility), the agent has more money to spend on treatment. As the insurer pays a fraction

$q \geq 0$ of these treatments, this leads to (weakly) higher costs for the insurer. Second, costs at full coverage ($q = 1$) do not vary in utility. Intuitively, if $q = 1$ treatments are for free for the agent and there is no reason to forgo treatments, irrespective of the level of $u^k \geq \bar{u}^k$.[41]

Because of (C1), the single crossing condition reads[42]

$$p_q(q, u^h, \theta^h) > p_q(q, u^l, \theta^l) > 0 \text{ for all } q \in [0, 1] \tag{SC}$$

and $u^h \geq \bar{u}^h$, $u^l \geq \bar{u}^l$ such that $p(q, u^h, \theta^h) = p(q, u^l, \theta^l)$. The intuition is the following. Suppose an indifference curve of type θ^h intersects with an indifference curve of type θ^l in some point (p, q). Then (SC) implies that the slope of the θ^h indifference curve will be higher. It follows that these two indifference curves can intersect only once.

We consider both the case where (SC) is satisfied and the case where it is violated (NSC). In both the SC and NSC cases, we maintain the assumption that $q = 1$ is the efficient insurance level (EI) for each type $\theta \in \{\theta^l, \theta^h\}$.

Assumption 3.3. *For a given utility level u^k, welfare (and therefore profits) are maximized at full coverage, i.e.*

$$\max_{q \in [0,1]} p(q, u^k, \theta^k) - c(q, u^k, \theta^k) \tag{EI}$$

is uniquely maximized by $q = 1$ for each $k \in \{h, l\}$ and $u^k \geq \bar{u}^k$.

This basically means that the insurance motive, i.e. transferring risk from a risk averse agent to a risk neutral insurer, is not overruled by other considerations. To illustrate, we do not assume that the low income agent's preference for health/treatment is so low that foregoing insurance would be socially optimal. Put differently, we assume that full insurance is socially desirable. Underinsurance–with no insurance as extreme case–results therefore not from first best but from informational distortions and price discrimination motives.

Our motivation for making this assumption is twofold. First, this assumption simply normalizes the socially efficient insurance level in the same way as in RS. Hence, we

[41] By assumption 3.3 full coverage is socially desirable. Hence we do not consider the case where insurance leads to inefficiency by inducing over-consumption of treatments.

[42] This is also called sorting, constant sign or Spence-Mirrlees condition (Fudenberg and Tirole, 1991, pp. 259).

only deviate from the RS set up by allowing for both SC and NSC. Second, we want to argue that under realistic assumptions, θ^h types have less than full insurance. If the optimal insurance level is actually below one, than this result would follow rather trivially. Another way of putting this is to say that a θ^h type would buy full insurance if he could choose from all actuarially fair insurance contracts. In this sense, the answer to our question "why healthy people have high coverage" is not simply that unhealthy people cannot afford actuarially fair insurance. In equilibrium, types are separated because the θ^h type prefers the cheap low coverage insurance above the expensive generous insurance contract.

To illustrate that the assumptions encompass standard models of the insurance literature, we show that the RS setup satisfies all of our assumptions.

Example 3.1. *In the RS setup an agent faces with probability θ a monetary loss D. He has initial wealth w and expected utility $u(q,p,\theta) = \theta v(w-(1-q)D-p)+(1-\theta)v(w-p)$ where $v' > 0$ and $v'' < 0$. Using the implicit function theorem, $p_q(1,u,\theta) = \theta D$ and therefore (C1) is satisfied. Note that (C1) will also be satisfied if w depends on θ. The insurer is risk neutral and has profits $p - \theta qD$. As profits do not depend on u (or w), assumption 3.2 is trivially satisfied. Since the agent is risk averse and the insurer is risk neutral, assumption 3.3 is also satisfied.*

3.2.2. Supply side

An insurer offers a menu of two contracts; one contract for each type. The contract of type θ^k consists of a coverage level q^k and a price p^k resulting in utility level u^k. The two contracts can be identical (pooling) or differ from each other (separating). In case of separating, the contracts have to satisfy the incentive compatibility (IC) constraints for each type:

$$p(q^l, u^l, \theta^l) \geq p(q^l, u^h, \theta^h) \qquad (IC_h)$$

$$p(q^h, u^h, \theta^h) \geq p(q^h, u^l, \theta^l) \qquad (IC_l)$$

The first constraint implies that the contract intended for θ^h (i.e. $(q^h, p(q^h, u^h, \theta^h))$) lies on a (weakly) lower indifference curve for θ^h than the contract that is meant for the θ^l type $(q^l, p(q^l, u^l, \theta^l))$. That is, the inequality implies $u(q^h, p^h, \theta^h) \geq u(q^l, p^l, \theta^h)$ where $p^i = p(q^i, u^i, \theta^i)$ with $i \in \{h, l\}$. This is illustrated in figure 3.1 where this inequality is

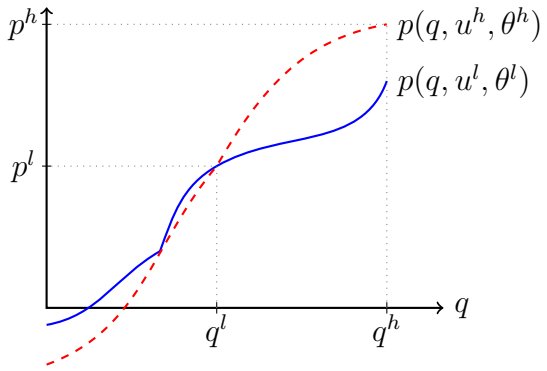


Figure 3.1: incentive compatibility constraints

binding: The θ^h indifference curve (dashed line) goes through both contracts. Similarly, the second inequality implies that $u(q^l, p^l, \theta^l) \geq u(q^h, p^h, \theta^l)$. In figure 3.1 ($IC_l$) is satisfied with inequality: The θ^l indifference curve through the (q^l, p^l) contract is below p^h at q^h.

Irrespective of the mode of competition and whether (SC) holds, we have the following result that we use below.

Lemma 3.1. *At least one type has full coverage. If the types are separated under the optimal contract scheme $(q^l, p^l), (q^h, p^h)$ with $q^l \neq q^h$, then at most one incentive constraint binds.*

Perfect competition

The literature on insurance models considers mostly perfect competition.[43] We show that with the assumptions made so far, perfect competition implies $q^h = 1$ (even if (SC) is not satisfied). Hence, in our model, market power on the insurance side is needed to get $q^h < 1$.

Following the RS definition of the perfect competition equilibrium, we require that (i) each offered contract makes nonnegative profits and (ii) given the equilibrium contracts there is no other contract yielding positive profits.

Proposition 3.1. *If an equilibrium exists under perfect competition, then $q^h = 1$.*

[43] See Jack (2006) and Olivella and Vera-Hernández (2007) for exceptions using a Hotelling model to formalize market power on the insurer side of the market. These papers assume that (SC) is satisfied and hence find efficient insurance for the θ^h type.

As is well known, existence of equilibrium in the RS framework is not guaranteed. Equilibrium does not exist if the only possible (separating) equilibrium is broken by a pooling contract. If the fraction of θ^h type agents in the population is high enough, then such a deviation to a pooling contract is not profitable and an equilibrium exists. If an equilibrium exists, it has $q^h = 1$.

The proposition shows that even with violations of single crossing, high risk types will get (weakly) higher coverage than low risk types. Hence we need to deviate from perfect competition to get $q^h < q^l$. Put differently, the positive correlation property holds in models of perfect competition irrespectively of single crossing. To explain violations of the positive correlation property which are pointed out in the empirical literature, it is necessary to deviate from the perfect competition assumption.

Indeed, recent research for the US (see Dafny (2010)) shows that health insurers have market power. More generally, in most countries where health insurance is provided by private companies, these firms tend to be big (due to economies of scale in risk diversification). Hence, one would expect them to have some market power.

Monopoly

Now, we want to consider an insurance monopolist. It turns out that the positive correlation property can be violated in the (NSC) case.

Proposition 3.2. *The type with the highest willingness to pay for full coverage, i.e. the type θ^k with highest $p(1, \bar{u}^k, \theta^k)$, obtains a full coverage contract in an insurance monopoly. Either his incentive compatibility or his individual rationality constraint is binding (or both). The other type's individual rationality constraint is binding.*

Let θ^k denote the type with the highest willingness to pay for full coverage. It follows from the proposition that θ^k obtains a contract $(q, p) = (1, p^k)$ for some $p^k \leq p(1, \bar{u}^k, \theta^k)$. The monopoly outcome is now pinned down by the choice of p^k. If $p^k = p(1, \bar{u}^k, \theta^k)$, then both individual rationality constraints are binding. In this case, type θ^{-k} might be excluded, i.e. θ^{-k} gets the contract $(0, 0)$. If $p^k = p(1, \bar{u}^{-k}, \theta^{-k})$, both types are pooled. If $p^k \in \langle p(1, \bar{u}^{-k}, \theta^{-k}), p(1, \bar{u}^k, \theta^k) \rangle$, the equilibrium separates the types and θ^{-k} gets an insurance contract with partial coverage. The optimal level of p^k is determined by the share of θ^k types in the population.

A direct implication of proposition 3.2 is that high risk types will always have full coverage if single crossing is satisfied. To see this, note that the indifference curve corresponding to $\bar{u}^k$ (that is the individual rationality constraint) goes through the origin $(p, q) = (0, 0)$ for both types. With (SC) the indifference curve of the high risk type is steeper and lies therefore above the individual rationality constraint of the low risk type for all coverage levels.

Without single crossing this is no longer the case. We will give a numerical example below where the low risk type θ^l has the higher willingness to pay for full coverage. Therefore, the low risk type will receive full coverage. If the types are separated (which depends on the share of each type in the population), we find that $q^h < q^l = 1$. Put differently, the positive correlation property no longer holds in the monopoly setup.

Of course, a monopolistic market structure is an extreme case and not entirely realistic in the health insurance market. However, the idea that the positive correlation property does not hold in the (NSC) case is more general. To illustrate this, we turn to an oligopoly setting next.

Oligopoly

This subsection uses a tractable duopoly model on the supply side. It serves as an illustration that the results from the monopoly setting also carry over to imperfect competition settings.

We assume that there are two profit maximizing insurers located at the end points 0 and 1 of a Hotelling line. Agents of both types are uniformly distributed over the $[0, 1]$ interval. The share of high risk types in the population is denoted by ϕ which is assumed to be independent from location $x \in [0, 1]$. An agent at position $x \in [0, 1]$ incurs transportation cost xt $((1 - x)t)$ when buying from insurer a (b). The agent maximizes the expected utility from the insurance contract minus the transportation costs. Each insurer offers a menu of contracts $\{(q^h, p^h), (q^l, p^l), (0, 0)\}$ where the first contract is intended for the θ^h type, the second for the θ^l type and the third "contract" denotes the agent's outside option of not buying insurance at all (which will not be used in equilibrium).[44] Insurers simultaneously offer menus and consumers choose their

[44]This means that "transportation costs" are not relevant for the participation decision. This ensures that firms compete also for high values of t. Hence, we rule out the case of (local) monopoly which was

preferred contract afterwards.

For the following result, we need the fairly standard assumption $u_{pp} \leq 0$, i.e. the higher the price the higher is the utility loss from a marginal price increase. Put differently, there is a decreasing marginal utility from other goods.

Proposition 3.3. *Assume* $u_{pp}(1, p, \theta^l) \leq 0$ *for* $p \in (0, p(1, \bar{u}^l, \theta^l))$. *If* $p(1, \bar{u}^l, \theta^l) > p(1, \bar{u}^h, \theta^h)$, *then there exist parameter values* $\phi > 0$ *and* $t > 0$ *such that type* θ^l *obtains a full coverage contract in a separating equilibrium.*

This proposition is similar to the result obtained for the monopoly setup. If the low risk type has a higher willingness to pay for full coverage, there are separating equilibria where he obtains full coverage. As noted above, since $p(0, \bar{u}^l, \theta^l) = p(0, \bar{u}^h, \theta^h)$, (SC) implies that $p(1, \bar{u}^l, \theta^l) < p(1, \bar{u}^h, \theta^h)$. But with a violation of single crossing it can indeed be the case that $p(1, \bar{u}^l, \theta^l) > p(1, \bar{u}^h, \theta^h)$. This is illustrated with a numerical example in section 3.4. Before going to the example, we first explain why single crossing is likely to fail in the health insurance context.

3.3. Income and health

We present a model where SC is violated due to differences in income between types. The idea that income differences can lead to a violation of single crossing can already be found in Wambach (2000), De Meza and Webb (2001) and in some sense also in Netzer and Scheuer (2010). The idea of these papers is that the degree of risk aversion depends on the wealth level. As the degree of risk aversion influences the shape of the indifference curve, this can lead to a violation of single crossing if the third derivative of the utility with respect to money has the corresponding sign. This effect exists in a simple RS framework that has no particular relation to the health insurance sector. As mentioned above, we are reluctant to make this assumption in a health insurance context: A violation of single crossing would only result if poor, high risk individuals are less risk averse than rich, low risk types. We do not believe that the uninsured, for instance in the US, forgo health insurance because they are risk neutral. Indeed, if this were the case for most of the uninsured, this would not call for government intervention.

analyzed above.

Also Fang et al. (2008) reject this channel as an explanation of advantageous selection in their empirical analysis of the medigap market.

In our model, there is an additional channel causing a violation of single crossing and consequently generating advantageous selection. This additional channel does not depend on third derivatives of the utility function. The idea of the model is that for $q < 1$ people have to finance a part of the costs of treatment out of their own pocket and low income agents may decide to choose cheaper treatment or forgo treatment altogether. This effect is documented in the medical literature, see for example Piette et al. (2004b), Piette et al. (2004a) or Goldman et al. (2007).[45] Put differently, the fact that health is a normal good can lead to a violation of single crossing. The reason is that poor, high risk types do not utilize the insurance fully when copayments are substantial. Therefore, their willingness to pay for a marginal increase in coverage can be lower than the one of rich, low risk types who fully utilize the insurance.

This utilization effect is well established in the medical literature. By extrapolating from their sample to the US population Piette et al. (2004a, p. 1786) conclude that "2.9 million of the 14.1 million American adults with asthma (20%) may be cutting back on their asthma medication because of cost pressures." They also document for a number of chronical conditions that people from low income groups are much more likely to report foregoing prescribed treatment due to costs.[46] Further examples can, for instance, be found in Piette et al. (2004b), Goldman et al. (2007), Schoen et al. (2010) or Schoen et al. (2008, pp. w305) who report that "[b]ased on a composite access indicator that included going without at least one of four needed medical care services, more than half of the underinsured and two-thirds of the uninsured reported cost-related access problems". With full coverage ($q = 1$) health insurance, such cost related access problems would not exist.

The utilization effect leads to a violation of single crossing if richer people face lower health risks, i.e. income and health risk are negatively correlated. This is also well

[45] While especially well documented in the health literature, this effect is not exclusively applicable to health insurance. Any insurance where the agent can choose between different ways to repair the damage ex post features this effect.

[46] For most chronic diseases people with income less than $ 20000 are roughly 2 (5) times more likely to forego prescribed treatment due to costs than people with an income between $ 20000 and $ 40000 (more than $ 60000); see table 3 in Piette et al. (2004a) for details.

documented in the empirical health literature, see for example Frijters et al. (2005), Finkelstein and McGarry (2006), Gravelle and Sutton (2009) or Munkin and Trivedi (2010). Potential explanations for this correlation between income and health include the following. High income people are better educated and hence know the importance of healthy food, exercise etc. Healthy food options tend to be more expensive and therefore better affordable to high income people. Or (with causality running in the other direction) healthy people are more productive and therefore earn higher incomes.

To illustrate how the described features of the health sector can lead to a violation of single crossing and also to exemplify the assumptions made in the reduced form model, we present a simple model of health insurance. We assume that a type θ consumer faces a health shock $s \in [0,1]$ with distribution (density) function $F(s|\theta)(f(s|\theta))$. We take $s = 1$ as the state in which the agent is healthy and needs no treatment. Lower health states s correspond to worse health. The assumption that the θ^h type has worse health than the θ^l type can now be stated as $F(s|\theta^h) > F(s|\theta^l)$ for each $s \in \langle 0,1\rangle$. In words, low s states are more likely for θ^h then for the θ^l type.

Once an agent receives a health shock $s < 1$, she can increase her health by treatment $h \in H(s)$ to health level $s+h$, where $H(s)$ denotes the set of possible treatments in state s. We assume that the set $H(s)$ is compact and $0 \leq s + h \leq 1$ for each $h \in H(s)$ and each $s \in [0,1]$. That is, treatment cannot lead to a higher health state than not falling ill. If $H(s)$ is a singleton, the consumer has no treatment choice. If the set $H(s)$ has more than one element, low income consumers with partial insurance, i.e. $q < 1$, may decide to choose cheaper treatment than if they have full insurance, i.e. $q = 1$.[47] We define $\bar{h}(s) = \max\{h \in H(s)\}$ as the best possible treatment and assume that $\bar{h}(s)$ is non-increasing in s. This means that a less afflicted agent (high $s < 1$) cannot increase his health by treatment more than an agent who is more seriously ill (low s). If $0 \in H(s)$, an agent can forgo treatment altogether.

[47]Implicitly, we assume that contracts cannot be contingent on treatment choice. Given the problems of verifiability of treatment and quantity choice as well as the possibility of doctor and patient to "collude" against the insurance (see Ma and McGuire (1997) for an analysis of these problems), this seems not unreasonable.

Let $w(\theta)$ denote the wealth (or income) of a type θ agent. Then we write

$$\begin{aligned} u(q,p,\theta) &= \int_0^1 \{v(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta))\}dF(s|\theta) \\ &\text{where } h(s,q,\theta) \text{ is defined as:} \\ h(s,q,\theta) &= \arg\max_{h\in H(s)} v(w(\theta)-p-(1-q)h, s+h) \end{aligned} \tag{3.2}$$

where $v(y,h)$ is the utility function of an agent which depends on consumption of other goods (y) and health (h). We assume that $v(y,h)$ satisfies $v_y, v_h > 0, v_{yy}, v_{hh} < 0$ and that health is a normal good: $v_{hy} \geq 0$. That is, utility increases in both health and consumption of other goods at a decreasing rate. As income increases, people's preference for health increases as well. In line with the empirical literature cited above, we assume that income and health status θ are negatively correlated: $w(\theta^h) \leq w(\theta^l)$.

Using this notation, we can write

$$c(q,u,\theta) = q\int_0^1 h(s,q,\theta)dF(s|\theta) \tag{3.3}$$

The first order condition for an interior solution $h(s,q,\theta) \in H(s)$ can be written as

$$(1-q)v_y(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta)) = v_h(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta)) \tag{3.4}$$

To see the implications of this model for single crossing, consider the slope of the indifference curves in (q,p)-space:

$$p_q(q,u,\theta) = -\frac{u_q}{u_p} = \frac{\int_0^1 v_y(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta))h(s,q,\theta)dF(s|\theta)}{\int_0^1 v_y(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta))dF(s|\theta)} \tag{3.5}$$

In words, the slope p_q equals the weighted average of treatment $h(s,q,\theta)$ over the states s with weight

$$\frac{v_y(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta))f(s|\theta)}{\int_0^1 v_y(w(\theta)-p-(1-q)h(s,q,\theta), s+h(s,q,\theta))dF(s|\theta)} \tag{3.6}$$

on state s (where the weights integrate to 1).

To illustrate (C1), assume that $s+\bar{h}(s) = 1$ (the best treatment makes a patient healthy again),[48] then it is routine to verify that

$$p_q(1,u,\theta) = \frac{\int_0^1 v_y(w(\theta)-p,1)\bar{h}(s)dF(s|\theta)}{\int_0^1 v_y(w(\theta)-p,1)dF(s|\theta)} = \int_0^1 (1-s)dF(s|\theta)$$

[48] Alternatively, we can assume that $\bar{h}'(s) \in \langle -1, 0\rangle$ such that $s+\bar{h}(s)$ is increasing in s. In words, if an agent falls ill, treatment does not bring back full health. Then a sufficient condition for (C1) is

where the last equality follows from the fact that $v_y(w(\theta) - p, 1)$ is constant in s. Note that we use here that $h(s, 1, \theta) = \bar{h}(s)$ for both types. If treatment is free ($q = 1$), each agent uses the highest treatment ($\bar{h}(s)$). The stochastic dominance assumption implies that θ^h puts more weight on low s states (where $\bar{h}(s) = 1 - s$ is high) compared to θ^l. Hence under these assumptions, condition (C1) is satisfied.

(SC) is satisfied if there are no wealth differences between types, i.e. $w(\theta^h) = w(\theta^l)$, and $H(s)$ satisfies some regularity condition. The idea is that without wealth differences, (3.4) yields for both types the same optimal treatment. Put differently, $h(s, q, \theta)$ is independent of θ. If patients choose more treatment in worse health states, single crossing will be satisfied: due to stochastic dominance, θ^h types have higher weight (3.6) on low s states with high $h(s)$. Hence, p_q in (3.5) is higher for θ^h than for θ^l types for all $q \in [0, 1]$. Treatment $h(s, q, \theta)$ is indeed non-increasing in s if $H(s)$ is well behaved: $H(s)$ is convex for each s and non-increasing in s.[49] It then follows from equation (3.4) –using the implicit function theorem– that

$$(-(1-q)^2 v_{yy} + 2(1-q)v_{yh} - v_{hh})\frac{dh}{ds} = v_{hh} - (1-q)v_{yh} \tag{3.7}$$

From the assumptions on v it follows that $h(s, q, \theta)$ is non-increasing in s. As $H(s)$ is non-increasing, this also holds true for boundary solutions where the implicit function theorem cannot be used.

However, if $w(\theta^h) < w(\theta^l)$ then $q < 1$ can imply that $h(s, q, \theta^h) < h(s, q, \theta^l)$. This follows from equation (3.4), since

$$\frac{dh}{dw} = \frac{-(1-q)v_{yy} + v_{hy}}{-(1-q)^2 v_{yy} + 2(1-q)v_{yh} - v_{hh}} > 0 \tag{3.8}$$

Hence, if $h(s, q, \theta^l) \in H(s)$ is an interior maximum, the θ^h type tends to choose lower treatment h. In words, since a fraction $1 - q$ of the treatment cost has to be paid by the insured, a low income θ^h patient may choose cheaper treatment than the richer θ^l type (as health is a normal good). Since he does not utilize the insurance as much as the (rich)

$v_{yyh} \geq 0$: Suppose it were the case that $\bar{h}'(s) = 0$, then $p_q(1, \cdot)$ would be the same for both types. $\bar{h}'(s) < 0$ will now put more weight on low states (as $\bar{h}(s_2) \leq \bar{h}(s_1)$ for $s_1 \leq s_2$). $v_{yyh} \geq 0$ guarantees that v_y is increasing less in s for the high type. Hence, putting more weight on low states where v_y is low affects the θ^l type less than the θ^h type. Consequently, (C1) is satisfied.

[49] We say that the set $H(s)$ is non-increasing in s if for each s_1, s_2 with $s_1 \leq s_2$ we have that for each $h \in H(s_2)$ there exists $h' \in H(s_1)$ such that $h' \geq h$. As a special case this includes the possibility that $H(s) = h(s)$ is a singleton, with $h'(s) < 0$.

low risk type, type θ^h has a lower marginal willingness to pay for insurance coverage (for q close to zero). However, for high levels of coverage, i.e. q close to 1, wealth differences matter less in the treatment choice because the patient does not have to pay (much) for the treatment. Consequently, although (C1) is satisfied with $w(\theta^h) < w(\theta^l)$, (SC) can be violated.

This model –where agents differ in income and treatment choice $h \in H(s)$ is endogenous– can generate the violation of (SC) mentioned above. In the following section, we give a numerical example where (SC) is indeed violated.

3.4. Example

As an example of an utility function that satisfies the assumptions (C1) and (EI) above and violates (SC), consider the following mean-variance utility set up.[50]

There are two states of the world: An agent either falls ill or stays healthy. The probability of falling ill is denoted by F^h ($F^l < F^h$) for type θ^h (θ^l). We choose $F^h = 0.07 > 0.05 = F^l$. Once an agent falls ill, the set of possible treatments is denoted by $H = \{\underline{h}, \bar{h}\}$. The utility of an agent of type $i = h, l$ with treatment choice $h \in \{\underline{h}, \bar{h}\}$ is written as:

$$\begin{aligned} u(q, p, \theta^i) = F^i(v(h, \theta^i) - (1-q)h) + (1-F^i)v(1, \theta^i) - p \\ -\tfrac{1}{2}r^i F^i(1-F^i)(v(1, \theta^i) - v(h, \theta^i) + (1-q)h)^2 \end{aligned} \tag{3.9}$$

where $v(h, \theta^i)$ denotes the utility for type $i = h, l$ of having health h and $r^i > 0$ denotes the degree of risk aversion. Hence an agent's utility is given by the expected utility minus $\frac{1}{2}r^i$ times the variance in the agent's utility. This is a simple way to capture that the agent is risk averse.[51]

[50] For the Python code used to generate this example, see:
`http://sites.google.com/site/janboonehomepage/home/webappendices`.

[51] When an agent of type i buys a product at price p that gives utility v, there are two ways to capture the marginal utility of income for agent i. First, overall utility can be written as $v - \alpha^i p$ where v is the same for each type i and α^i can differ. Low income types are then modelled to have high α^i; high marginal utility of income. Alternatively, one can write $v^i - \alpha p$ where α is the same for all types. Then low income types have low v^i. We have chosen the latter formalization with $\alpha = 1$. The assumption

Along an indifference curve where u is fixed, we find the following slope:

$$\frac{dp}{dq} = F^i h(q,\theta^i) + r^i F^i(1-F^i)(v(1,\theta^i) - v(h(q,\theta^i),\theta^i) + (1-q)h(q,\theta^i))h(q,\theta^i) \quad (3.10)$$

where $h(q,\theta^i)$ is the solution for h solving

$$\max_{h\in\{\underline{h},\bar{h}\}} v(h,\theta^i) - (1-q)h.$$

In words, once an agent falls ill, she decides which treatment to choose based on the benefit $v(h,\theta^i)$ and the out-of-pocket expenses $(1-q)h$.

With the parameter values that we consider below, it is the case that

$$r^h F^h(1-F^h)(v(1,\theta^h) - v(\bar{h},\theta^h))\bar{h} = r^l F^l(1-F^l)(v(1,\theta^l) - v(\bar{h},\theta^l))\bar{h} \quad (3.11)$$

In words, at $q=1$ (where both types choose the highest treatment $\bar{h}$) the variance terms in the slope dp/dq (equation (3.10)) are equalized. Hence, assumption (C1) is satisfied because $F^h\bar{h} > F^l\bar{h}$ in equation (3.10).[52]

For the numerical example, we assume $\bar{h} = 0.6, \underline{h} = 0.2$ and the associated utilities for the θ^h type equal $v(1,\theta^h) = 0.9, v(\bar{h},\theta^h) = 0.7, v(\underline{h},\theta^h) = 0.45$ and similarly for the θ^l type: $v(1,\theta^l) = 1.1, v(\bar{h},\theta^l) = 0.9, v(\underline{h},\theta^l) = 0.5$. Hence, having high health is more important for the θ^l type compared to the θ^h type. This implies that θ^l type is willing to spend more on treatment than the θ^h type. Since $0.9-0.6 \geq 0.5-0.2$ the θ^l type chooses $\bar{h}$ even if $q=0$ (and the inequality is strict for $q>0$). This implies that condition (EI) is satisfied for the θ^l type as q does not affect treatment choice and higher q leads to more insurance (provided by a risk neutral insurer). The θ^h type chooses $\bar{h}$ if $q=1$ but prefers $\underline{h}$ for low values of q. In particular, for $q=0$ we have $v(\bar{h},\theta^h) - \bar{h} < v(\underline{h},\theta^h) - \underline{h}$. Let $\tilde{q}$ denote the value for q such that the θ^h type is indifferent between treatment $\bar{h}$ and $\underline{h}$:

$$v(\bar{h},\theta^h) - (1-\tilde{q})\bar{h} = v(\underline{h},\theta^h) - (1-\tilde{q})\underline{h} \quad (3.12)$$

that treatment is a normal good is then implemented by assuming that

$$v(\bar{h},\theta^h) - v(\underline{h},\theta^h) < v(\bar{h},\theta^l) - v(\underline{h},\theta^l).$$

[52]Note also that the numerical values below are chosen such that the variance term in the utility function is also equal at full coverage (and weakly higher for θ^h if $q<1$). Therefore, the violation of single crossing in our example is due to the different utilization of health insurance and not to differences in risk aversion that were the driving force in other papers on the violation of single crossing.

To verify that (EI) is satisfied for the θ^h type, we proceed in two steps. First, consider $q > \tilde{q}$ such that $h(q, \theta^h) = \bar{h}$. Then increasing coverage q reduces the variance in utility for the risk averse θ^h type and hence (EI) is satisfied for $q > \tilde{q}$. Now consider $q < \tilde{q}$ such that $h(q, \theta^h) = \underline{h}$. In order to satisfy (EI), it must be the case that profits (price[53] minus expected costs) when offering full coverage are higher than profits when offering a partial coverage contract yielding the same utility. This can be written as:

$$F^h(v(\bar{h}, \theta^h) - \bar{h}) + (1 - F^h)v(1, \theta^h) - u^h - \tfrac{1}{2}r^h F^h(1 - F^h)(v(1, \theta^h) - v(\bar{h}, \theta^h)^2 \geq$$
$$F^h(v(\underline{h}, \theta^h) - \underline{h}) + (1 - F^h)v(1, \theta^h) - u^h - \tfrac{1}{2}r^h F^h(1 - F^h)(v(1, \theta^h) - v(\underline{h}, \theta^h) + (1 - q)\underline{h})^2$$

Note that the right hand side of this inequality increases in q and hence is highest at $\tilde{q}$. In our numerical example, we choose r^h such that the inequality holds with equality at $q = \tilde{q}$.[54] This implies that it is satisfied for all $q \leq \tilde{q}$ and hence (EI) is satisfied.

With the parameter values above, it is routine to verify that (SC) is violated. Figure 3.2 shows two indifference curves for the θ^l type (in red) and one for the θ^h type (in blue). Indeed, for $q < \tilde{q}$ the indifference curve for the θ^l type is steeper than for the θ^h type. This is due to the fact that the θ^l type buys the expensive treatment $\bar{h}$ while the θ^h type buys $\underline{h}$. The kink in the indifference curve for the θ^h type happens at $\tilde{q}$ where the θ^h type switches from the cheap to the more expensive treatment. Hence small increases in q for $q > \tilde{q}$ are worth more to the θ^h type than small increases in $q < \tilde{q}$. In fact, the figure shows that for $q > \tilde{q}$, the indifference curve for the θ^h type is steeper than the one for the θ^l type. This is the violation in single crossing.

Hence in a simple mean-variance utility framework, it is straightforward and intuitive to generate a violation of (SC).

In our numerical example, the low risk type has a higher willingness to pay for full coverage compared to the high risk type. This can be seen in figure 3.2: The solid lines are the indifference curves of both types that go through the empty contract $(0, 0)$, i.e. the indifference curves corresponding to individual rationality. Willingness to pay for full coverage is given by the value at $q = 1$ where the low risk type's indifference curve is above the high risk type's indifference curve.

By proposition 3.2, this implies that in a monopoly framework $q^l = 1$. If the share of

[53] Price is derived from solving equation (3.9) for p^h as a function of u^h and $q^h < \tilde{q}$.

[54] Given this value of r^h, r^l is chosen to satisfy equation (3.11).

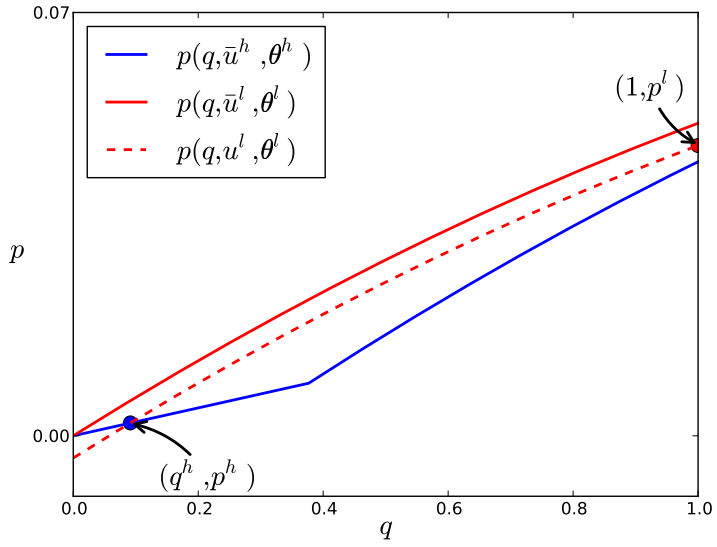


Figure 3.2: Example with parameter values in section 3.4 and $t = 0.018$

low risk type's is high enough, the optimal monopoly menu separates the two types and the positive correlation property will be violated.

3.4.1. Duopoly

In this subsection, we use the example to illustrate the logic behind proposition 3.3. We will show that it is straightforward to find examples where $q^l = 1$ and $q^h < 1$. The easiest way to do this is to find parameter values such that the individual rationality (IR) curve (that is, the indifference curve $p(q, \bar{u}, \theta)$) for the θ^l type lies everywhere above the IR curve for the θ^h type. As shown in figure 3.2, this is the case for the parameter values of our numerical example. Clearly, the Hotelling equilibrium contracts have to lie on or below the relevant IR curves.

First, assume that $\phi = 0$. In words, there are only θ^l types. Then it is routine to verify that $q^l = 1$ (because of assumption 3.3) and the Hotelling equilibrium price on the θ^l-market equals $p^l = F^l\bar{h} + t$.[55] This contract is denoted $(1, p^l)$ in figure 3.2 for the parameter values given above and $t = 0.018$. As this contract lies below θ^l's IR curve, it is, indeed, the equilibrium outcome. Let $u^l_{hotel.}$ denote θ^l's utility level associated with the

[55]Recall that in a Hotelling model with constant marginal costs c, the equilibrium price is given by $c + t$. See, for instance, Tirole (1988, pp. 280) or the proof of proposition 3.3 in the appendix with $u_p = -1$.

$(1, p^l)$ contract: $u^l_{hotel.} = u(1, p^l, \theta^l)$. Contract (q^h, p^h) (although not bought by anyone as $\phi = 0$) is defined by the intersection of indifference curve $p(q, u^l_{hotel.}, \theta^l)$ (dashed curve in the figure) and θ^h's IR curve. This is the best contract on θ^h's IR curve that satisfies θ^l's incentive compatibility constraint.

Now increase ϕ slightly to $\phi > 0$ (but small). We claim that this results in an equilibrium with $q^l = 1 > q^h$. For this to be an equilibrium we need that the indifference curve for the θ^l type at $q = 1$ lies above the indifference curve for the θ^h type at $q = 1$. Note that the equilibrium indifference curve for the θ^h type $(p(q, u^h_{hotel}, \theta^h))$ cannot lie above θ^h's IR curve. Hence, a sufficient condition for an equilibrium with $q^l = 1 > q^h$ is that θ^l's indifference curve $p(q, u^l_{hotel.}, \theta^l)$ at the new Hotelling equilibrium lies above θ^h's IR curve at $q = 1$. This is formally shown in the proof of proposition 3.3 and is intuitively clear: Small changes in ϕ will lead to small changes in the indifference curve $p(q, u^l_{hotel.}, \theta^l)$. As this curve is above θ^h's IR curve at $q = 1$ in case $\phi = 0$, it will be above θ^h's IR curve for small positive values of ϕ.

Hence a straightforward way to generate equilibria where the positive correlation property fails, is to find examples where the IR constraint for the θ^l type lies above the IR constraint for the θ^h type for each $q \in \langle 0, 1]$. Then there exist $t > 0$ and $\phi > 0$ such that the example has an equilibrium with $q^l > q^h$.

3.5. Conclusion

Standard insurance models, e.g. Rothschild and Stiglitz (1976) or Stiglitz (1977), predict higher coverage for agents with higher risks. We show that this prediction no longer holds if single crossing is violated and firms have market power.

In the health care sector, agents with higher income have lower risks and more insurance. Put differently, the predictions of the standard insurance model with single crossing are contradicted by the data. We show that the negative correlation between income and risk can cause a violation of single crossing. With a violation of single crossing, the empirical findings in the health literature can be reconciled with a standard insurance model.

From an empirical point of view, our paper casts doubt on the positive correlation

test: Given our result that separating equilibria exist in which agents with higher risk have less coverage (negative correlation), it is evident that the results of such a test have to be interpreted with care. In particular such a test cannot be used to test for the presence of asymmetric information when single crossing is violated.

Further, our analysis shows that one should be careful with interpreting actual expenditure as a signal of risk type. Indeed, in equilibria in which the low risk type has full coverage it is not clear which type has the highest expected health care expenditure. On the one hand, the high risk type is more likely to fall ill and need treatment. On the other hand, because the low type has a more generous contract, conditional on falling ill the low type spends more on treatment than the high type (who faces high co-payments). Expected costs of a type are the product of the former and the latter and cannot be ordered unambiguously. Hence, it is not necessarily correct to identify types based on their expenditure.

We conclude with a discussion of advantageous selection. The standard in this literature is to assume that people differ in their preferences for risks. If high risk individuals are less risk averse than low risk people, it can happen that consumers who are willing to pay the most for health insurance are people with low expected health care costs. Hence, offering health insurance with high coverage is especially attractive for agents with low expected costs: advantageous selection. The implication of some advantageous selection models is that policies that stimulate insurance coverage are welfare reducing, see for example De Meza and Webb (2001). In fact, there may be over-insurance in equilibrium. See Einav and Finkelstein (2011) for a recent review of advantageous selection and empirical papers documenting this in health care markets.

In our model, we also see that at the margin low risk types are willing to pay more for insurance than high risk types. This is caused by the fact that at less than perfect coverage, low income, high risk types tend to reduce expenditure on treatments. Basically, they cannot afford the treatments that they need. Hence, although the equilibrium is an advantageous selection equilibrium, in our model stimulating insurance coverage (e.g. through mandatory insurance at full coverage) is efficient (because of assumption 3.3).

3.6. Appendix: Proofs

Proof of lemma 3.1. We start with the proof of the second statement. Suppose both incentive constraints were binding, i.e. θ^h and θ^l are both indifferent between the two contracts. First, look at the case where $q^h, q^l < 1$. Call the utility levels of the two types under the equilibrium contracts u^l and u^h. Now take the indifference curves corresponding to these utility levels and call them $p(q, u^l, \theta^l)$ and $p(q, u^h, \theta^h)$ and define $\iota = \arg\max_{k\in\{l,h\}} p(1, u^k, \theta^k)$. Changing θ^ι's menu point to $(1, p(1, u^\iota, \theta^\iota))$ will increase profits by assumption 3.3. By the definition of ι, this change is also incentive compatible.

Second, take the case where $q^k = 1$ and $q^{-k} < 1$ for some $k \in h, l$ and suppose again that both incentive constraints were binding. But according to assumption 3.3 pooling on the contract of θ^k would lead to higher profits. Hence, at most one incentive constraint is binding.

$q^\iota = 1$ follows from the argument in the first step and therefore at least one type has to have full coverage. *Q.E.D.*

Proof of proposition 3.1. Suppose to the contrary that $q^h < 1$ in equilibrium. Lemma 3.1 implies then that $q^l = 1$. Note that θ^h has to prefer his contract strictly to the θ^l contract: Otherwise, pooling on the θ^l contract would be a profitable deviation by assumption 3.3. Given that (IC_h) is not binding, the θ^l contract leads to zero profits: Otherwise, marginally decreasing its price (and thereby attracting all demand of θ^l types) would be a profitable deviation.

The contract (q^h, p^h) leads to nonnegative profits; otherwise it would not be offered in equilibrium.[56] Denote by u^h the utility level θ^h derives from (q^h, p^h) and by $p(q, u^h, \theta^h)$ the indifference curve of θ^h associated with his contract. By assumption 3.3, the contract $(1, p(1, u^h, \theta^h))$ for type θ^h yields higher profits than (q^h, p^h). For $\varepsilon > 0$ small enough, the contract $(1, p(1, u^h, \theta^h) - \varepsilon)$ is strictly preferred by θ^h to (q^h, p^h) and yields higher profits than (q^h, p^h). If the contract $(1, p(1, u^h, \theta^h) - \varepsilon)$ also attracts θ^l types, profits from those θ^l types will be positive as well as those are better risks. This would be an additional gain as it was shown above that the θ^l contract yields zero profits. Therefore, $(1, p(1, u^h, \theta^h) - \varepsilon)$ is a profitable deviation, i.e. a contract with strictly positive profits and demand. Consequently, $q^h < 1$ cannot be an equilibrium. *Q.E.D.*

[56] In fact, it has to be a zero profit contract.

Proof of proposition 3.2. Define $\iota = \arg\max_{k\in\{h,l\}} p(1, \bar{u}^k, \theta^k)$. By lemma 3.1, one type has full coverage. Suppose that $q^\iota < 1$ and therefore $q^\kappa = 1$ with $\kappa \in \{h, l\}$ and $\kappa \neq \iota$. Note that the individual rationality constraint of θ^ι cannot be binding as otherwise θ^ι would misrepresent as θ^κ by the definition of ι. But then the incentive compatibility constraint of θ^ι has to be binding as the monopolist could increase p^ι otherwise. By assumption 3.3, the monopolist could achieve a higher profit by pooling both types on θ^κ's contract. This contradicts the optimality of $q^\iota < 1$.

If both types are pooled, the optimal contract will be $(q, p) = (1, p(1, \bar{u}^\kappa, \theta^\kappa))$ and the individual rationality constraint of θ^κ will be binding. If the types are separated, the incentive compatibility constraint of θ^κ cannot bind: Since $q^\iota = 1$, pooling on θ^ι's contract would lead to higher profits by assumption 3.3 if the incentive constraint was binding. As increasing p^κ relaxes the incentive compatibility constraint of θ^ι, the individual rationality constraint of θ^κ has to bind: Otherwise, increasing p^κ would increase profits.

Last note that increasing p^ι would be feasible and increase profits if neither the incentive compatibility nor the individual rationality constraint of θ^ι was binding. *Q.E.D.*

Proof of proposition 3.3. The first step is to analyze the game where $\phi = 0$, i.e. a standard Hotelling game where only low risk type exist. From assumption 3.3, $q^l = 1$ in this setting and firms only compete in prices. By assumption 3.2, costs do then not depend on price and can be denoted by $\bar{c}$. A firm maximizes $(p^l - \bar{c})(\frac{1}{2} + \frac{u(1,p,\theta^l) - u^b}{2t})$ where u^b is the utility offered by the other firm. Because of the assumption $u_{pp}(1, p, \theta^l) \leq 0$, the objective is concave and the best response is defined by the first order condition

$$t + u(1, p, \theta^l) - u^b + (p - \bar{c})u_p(1, p, \theta^l) = 0.$$

Note that there is a symmetric equilibrium defined by the equation $(p-\bar{c})u_p(1, p, \theta^l) = -t$. The left hand side of this equation is decreasing in p and therefore there is only one symmetric equilibrium. We will now argue that there are also no asymmetric equilibria, i.e. the game has a unique equilibrium. The argument is that the slope of the best response function is less than one whenever crossing the 45° line where $p = p^b$: By the implicit function theorem, $p'(p^b) = \frac{-u_p(1,p^b,\theta^l)}{-2u_p(1,p,\theta^l)+(p-\bar{c})u_{pp}(1,p,\theta^l)}$. Consequently, $0 < p'(p^b) < 1$ whenever $p = p^b$. Given that there is a symmetric equilibrium where $p = p^b$, this implies that the best response functions can only intersect once, i.e. there is a unique equilibrium.

The second step is to see that by choosing t appropriately the game with $\phi = 0$ leads to an equilibrium price $p^* \in (p(1, \bar{u}^h, \theta^h), p(1, \bar{u}^l, \theta^l))$. This simply follows from the fact that t "shifts" the first order condition above. For the rest of the proof let t take such a value.

The third step is to show that for ϕ small enough $q^l = 1$ and $q^h < 1$.[57] By lemma 3.1, at least one type has to have full coverage. If $q^h = 1$, then $p^h \leq p(1, \bar{u}^h, \theta^h)$ to satisfy individual rationality. Suppose, both types were pooled. Recall that $p^l < p(1, \bar{u}^h, \theta^h)$ is not an equilibrium when $\phi = 0$, i.e. there exists a profitable deviation for at least one insurer. For ϕ small enough, this deviation is again profitable as profit functions are continuous in all variables and parameters. Hence, there cannot be pooling for small but positive ϕ. Next suppose there were separating equilibria with $q^h = 1$ and $q^l < 1$ for all $\phi > 0$. By assumption 3.3, q^l has to converge to 1 as ϕ decreases (otherwise setting $q^l = 1$ and adjusting the price to keep u^l fixed is a profitable deviation for small enough ϕ). But then the same argument as in the pooling case shows that there is a profitable deviation for ϕ small enough. It follows that $q^l = 1$ and $q^h < 1$ for small enough $\phi > 0$. *Q.E.D.*

[57] Existence of equilibrium follows from Glicksberg (1952).

Chapter 4

Adverse selection without single crossing

4.1. Introduction

Adverse selection models–sometimes also referred to as screening models–are among the most used microeconomic models. The main feature of these models is that one (or more) agents have private information which is relevant for transactions with other players. This private information can be the efficiency of a firm in models of regulation (Baron and Myerson, 1982; Laffont and Tirole, 1987), the productivity of a worker in labor market (Guasch and Weiss, 1981) as well as in optimal taxation models (Mirrlees, 1971), the risk of an accident in insurance models (Stiglitz, 1977) or the willingness to pay for a product in models of monopoly pricing (Mussa and Rosen, 1978) and auctions (Myerson, 1981).

In screening models, the principal offers a menu of options from which the agent who has private information will choose his preferred option. The chosen option will normally not be what a benevolent planner with complete information would assign. Hence, informational distortions exist and will reduce welfare. The reason in a nutshell is that the agent reveals (some of) his private information by his choice. This will not be costless for the principal who designs the menu: The agent receives an informational rent. By distorting the menu away from first best, the principal can reduce this informational rent to his own benefit.

In the regulation example, a regulator will want a more efficient firm to produce a higher quantity than a less efficient firm. But an efficient firm could claim to be inefficient and choose the (low quantity) option intended for an inefficient firm from the menu. By distorting the quantity intended for an inefficient firm, the regulator can make such misrepresentation less attractive for an efficient firm. Consequently, an efficient firm is

willing to choose the high quantity option at a lower price and the regulator can save money.

Single crossing–which is also referred to as Spence-Mirrlees condition or sorting condition–is a technical assumption usually made in adverse selection models. In one dimensional models, single crossing states that types[58] can be ordered according to their marginal rate of substitution between monetary transfers and the decision, e.g. produced quantity in the regulation example above. With the usual quasilinear preferences, single crossing is equivalent to a type ordering according to marginal utilities.

In the regulation example above, the firm's cost depends on quantity and type. Single crossing means that higher types have lower marginal costs for any admissible quantity. Single crossing is violated if such an ordering is impossible, e.g. a higher type has lower marginal costs for high quantities but higher marginal costs for low quantities.

This paper analyzes a screening model in which single crossing is violated. Agents have quasilinear preferences and a one-dimensional type. The setting allows for a one time violation of single crossing; e.g. for a given quantity, marginal costs are first increasing and then decreasing in type. Without single crossing, local incentive compatibility does no longer guarantee global incentive compatibility. Therefore, non-local incentive compatibility constraints have to be taken into account. The paper analyzes monotone solutions in this setup, e.g. situations in which higher types produce higher quantities under the optimal contract. Sufficient conditions for the existence of a monotone solution and an algorithm to calculate such a solution are presented.

With single crossing, there is no distortion at the top and the distortion for all types goes in the same direction, e.g. all types produce a quantity which is weakly below their first best quantity. If single crossing is violated, neither result has to hold. The reason is that binding non-local incentive constraints will counteract the normal distortion stemming from local incentive compatibility and rent extraction motives. A rough intuition for this result is the following: With single crossing, distortions occur because the principal wants to lower the agent's informational rent. If a non-local incentive constraint is violated, a certain type's rent at "his contract" is too low compared with another type's contract. To satisfy his non-local incentive constraint, his rent has to be increased. Re-

[58] A "type" is an agent with a specific private information attribute, see Harsanyi (1967). In the regulation example types correspond to cost functions of the firm.

ducing the normal distortion (or even distorting the decision in the opposite direction) will result in such an increase.

The following section gives several examples of settings in which single crossing is violated. The related literature is reviewed in section 4.3 and the formal model is introduced in section 4.4. Section 4.4 also states a sufficient condition for the existence of a monotone solution. Section 4.5 analyzes the solution: Subsection 4.5.1 introduces necessary conditions which have to hold at types where non-local incentive constraints are binding. The core of the paper are the subsections 4.5.2 and 4.5.3: The former characterizes monotone solutions while the latter focuses on the special case of monotone and continuous solutions. An explanation why the no-distortion-at-the-top property is not always satisfied follows in subsection 4.5.4. Section 4.6 discusses some assumptions and section 4.7 concludes by pointing out direct implications of the paper for applied work.

4.2. Examples

This section illustrates why single crossing is violated in a number of reasonable economic settings. Section 4.2.1 gives several examples with a common theme: There are more than one input/option/relevant characteristic. It is then a priori unclear (and sometimes even unreasonable) that a higher type is "better" on all dimensions. But this is exactly what single crossing would require. Section 4.2.2 presents a three type example which shows not only how preferences can directly violate single crossing but also gives some intuition for the results of the paper.

4.2.1. Example settings where single crossing is violated

Example 1: two factor production. Take a setting where a firm or government has to contract with the provider of a good (input or public good/infrastructure etc.). If the principal is a government, this setting is mathematically equivalent to incentive regulation (compare for example Laffont and Tirole (1993)). Assume now that production uses variable input factors in fixed proportions. These input factors fall in one of two groups depending on how they affect costs: The first group are inputs which increase costs proportional to output, e.g. energy costs and unskilled labor. The second group are inputs increasing costs convexly in output, e.g. skilled labor (due to search costs)

and machine utilization. Type indexes the possible production technologies and denotes which of these two groups of inputs is used more efficiently by the firm. A cost function representing this setting could be[59]

$$c(q,\theta) = \theta q + \frac{q^2}{\theta} + \gamma(\theta)$$

where $\gamma(\theta)$ are (possibly type dependent) fixed costs. To give a more straightforward interpretation, let type represent whether a firm uses a labor intensive or capital intensive production technology. A labor intensive production technology requires especially unskilled labor which can be hired at a constant market wage (linear part). A capital intensive technology requires less but more skilled employees. Finding them is increasingly difficult and results therefore in convexly increasing costs. A more capital intensive technology might be associated with higher fixed costs (of capital).

Whether marginal costs $c_q(q,\theta) = \theta + \frac{2q}{\theta}$ are increasing or decreasing in type depends on the produced quantity q. Put differently, the cross partial derivative $c_{q\theta}(q,\theta) = 1 - 2q/\theta^2$ can change sign and therefore single crossing is violated. The idea is simple: For low quantities, the linear part of the cost function dominates marginal costs and therefore high types have higher marginal costs. For high quantities, the convex part of the cost function is more relevant and therefore high types have lower marginal costs.

It should be mentioned that the cost function in this example can be viewed as a simplified version of the flexible fixed cost quadratic cost function suggested by Baumol et al. (1982). Beard et al. (1991) estimate such a cost function for savings and loans associations. Interestingly, they allow for two unobservable types of production technology in their estimation. In table 5, Beard et al. (1991) report estimated costs for the two types ("mixtures" in their language) at different quantity levels. If one interprets estimated cost differences between the output levels as marginal costs, it turns out that mixture 1 has lower marginal costs at low output levels but higher marginal costs at high output levels. Hence, single crossing is violated.

Example 2: hiring talent and productivity. This example is in the context of compensation of workers.[60] The principal is the owner of a firm and the agent a worker the firm wants to hire. For the quality of the worker talent and effort are relevant, e.g. talent is what the worker produces in a regular working time like the 40 hours week and

[59] The alternative cost function $c(q,\theta) = \theta q + (1-\theta)q^2 + \gamma(\theta)$ also violates single crossing.

[60] A similar example can be found in Araujo and Moreira (2010).

effort is the additional time he is willing to invest. Assume the worker creates value $q = e\theta + T$ where T is his talent, e is the unobservable effort and θ is his type. The owner of the firm observes a public signal, e.g. education, which is a mix of talent and productivity (he does not observe T and θ directly). To be precise, assume that the signal is $\sigma = \theta * T$. Given this signal, a more productive worker will have lower talent and vice versa. The production function of the manager for a given signal is $q = e\theta + \sigma/\theta$ where q is the quantity/value produced by the worker. If costs of effort are e^2 and the worker's preferences are quasilinear in money, his utility function can be written as

$$u(q, \theta) = w - \frac{(q - \sigma/\theta)^2}{\theta^2} \tag{4.1}$$

where w is wage. It is easy to check that single crossing is violated. The intuition is that a low type can produce a low output q without much effort just within the regular working time. Hence, his marginal costs of effort (and therefore of q) are low. A high type already has to exert some effort to reach the same output level and therefore his marginal costs of effort (and q) are higher. Note here that the contract is conditional on education, i.e. given σ a more productive type will be less talented. For high output, where effort of both types is substantial, higher types have lower marginal costs since they are more productive.

Example 3: common agency. As already mentioned in Martimort and Stole (2009), violations of single crossing can arise if more than one principal contract with the same agent. Interestingly, the utility function itself will satisfy single crossing (for a fixed decision with the other principal) and the violation of single crossing results from the existence of multiple principals. This example tries to convey the idea in a simplified setup.

The source of hidden information in this example is the inability of firms to know the exact preferences of a customer. A firm cannot observe the preferences of a customer but it can engage in non-linear pricing, i.e. second degree price discrimination.

Say, consumers can buy two goods which are imperfect substitutes: Good A is sold only by firm A while good B is available on a perfectly competitive market at a constant per unit price p^B.[61] For concreteness, let the demand function for good B of a type θ

[61] See Martimort and Stole (2009) for a model in which the second good is also offered by a strategically acting principal.

consumer be

$$q^B(q^A, \theta) = \theta(\beta - p^B - \delta q^A) \tag{4.2}$$

which means that type rotates the inverse demand function outwards. The following quadratic utility function yields such a demand function:

$$u(q^A, q^B, \theta) = \alpha q^A + \beta q^B - \frac{\gamma}{2\theta}(q^A)^2 - \frac{1}{2\theta}(q^B)^2 - \delta q^A q^B - p^B q^B - p^A q^A$$

Firm A faces consumers buying product B according to (4.2). By plugging (4.2) into the utility function, one can obtain utility as a function of q^A and θ alone, i.e. $v(q^A, \theta) = u(q^A, q^B(q^A, \theta), \theta)$. This is the utility function firm A has to take into account in its profit maximization problem. Because consumers buy also product B, single crossing is violated:

$$\begin{aligned} v_{q^A\theta}(q^A, \theta) &= u_{q^A\theta}(q^A, q^B(q^A, \theta), \theta) + u_{q^B\theta}(q^A, q^B(q^A, \theta), \theta)\frac{\partial q^B(q^A, \theta)}{\partial q^A} \\ &= q^A\left(\frac{\gamma}{\theta^2} + \delta^2\right) - \delta(\beta - p^B) \end{aligned}$$

Clearly, $v_{q^A\theta}$ is negative for low q^A and positive for high q^A. The reason for the violation of single crossing is that high type consumers have, on the one hand, a higher marginal willingness to pay because of their high type (that is the $\frac{\gamma}{2\theta}(q^A)^2$ term in the utility function $u(q^A, q^B, \theta)$). On the other hand, a high type buys more of product B which reduces his willingness to pay for product A as the two goods are substitutes.

The basic intuition of this example is also reflected in the following story: Think of fixed line internet access. Heavy internet users will certainly have a higher marginal utility from the fifth gigabyte of data than light users. If heavy users, however, also own smartphones with internet access (and light users do not), light users will probably have a higher willingness to pay for the first 50 megabyte: They cannot switch to their mobile devices to check emails etc.. Hence, single crossing would be violated.

Example 4: insurance with mean variance utility. An agent faces a risk of losing a (money equivalent) amount D with probability θ where θ is private information. His preferences are given by the mean variance utility function

$$\begin{aligned} u(q, \theta) &= \mathbb{E}[wealth] - 1/2r\, Var[wealth] \\ &= \theta(w - (1-q)D) + (1-\theta)w - p - 1/2r\theta(1-\theta)(1-q)^2D^2 \end{aligned}$$

where p is the insurance premium of an insurance covering fraction q of the loss, w is initial wealth and $r > 0$ is a measure of risk aversion. The cross derivative $u_{q\theta} =$

$D + (1-q)rD^2(1-2\theta)$. If $\theta > 1/2$ and $rD > 1$, the cross derivative can change sign depending on q. Hence, single crossing is violated.

The intuition is that for $\theta > 1/2$ a higher risk also implies less variance. Consequently, a higher type is on the one hand more eager to buy insurance because he has a higher risk on the other hand he is less eager to buy insurance because there is less variance in his payoffs. At full coverage, i.e. for $q = 1$, the payoff variance is zero and the latter effect is no longer present. For lower coverage levels, however, it might dominate.

Example 5: nonlinear pricing with heterogeneity in demand elasticity. A monopolist sets a nonlinear price schedule. Consumers differ in their demand elasticity. More specifically, their utility from consuming q units for a price p is given by

$$u(q, \theta) - p = q^{\theta} - p$$

where θ is distributed on a subset of $(0, 1)$. If the relevant quantities are in $(0, 1)$, higher types have a lower willingness to pay and a higher price elasticity of demand.[62] The cross-derivative $u_{q\theta} = q^{\theta-1}(1 + \theta ln(q))$ changes sign at $q = e^{-1/\theta}$. Hence, single crossing is violated.

4.2.2. Three type example

In the airline industry there are often three classes: First class, business class and economy class. The simplest model leading to this result is a model of non-linear pricing with three possible consumer types which differ in their taste for quality. For economy class, think of poor leisure traveler with a low willingness to pay for quality, say $\theta^l q$ where q is quality and θ^l is some positive number. For the first class, think of luxury travelers with a high willingness to pay for quality, say $\theta^h q$ with $\theta^h > \theta^l$. So far, single crossing is satisfied: Luxury travelers have a higher marginal willingness to pay for quality than poor leisure travelers at every quality level.

The third group of travelers are business travelers and I define them the following way: Business travelers have a very high willingness to pay for the first units of quality, e.g. for a higher seat pitch, a socket at the seat, internet access and the option to leave the plane first. However, they have a lower willingness to pay than the luxury traveler

[62] The price elasticity of demand is here defined as the relative demand change caused by a 1% increase in the marginal price. Using the first order condition $p'(q) = \theta q^{\theta-1}$, the elasticity can be derived as $|1/(\theta - 1)|$.

at high levels of quality, e.g. for an exquisite wine card, limousine service at the airport or a personal flight attendant.[63] Their preferences can therefore be represented by

$$v^b(q) = \begin{cases} \theta^{max} q, & \text{for } q \leq \tilde{q} \\ \theta^{max}\tilde{q} + (q - \tilde{q})\theta^b, & \text{for } q > \tilde{q} \end{cases}$$

with $\theta^b < \theta^h < \theta^{max}$. Now single crossing is violated since business travelers have a higher willingness to pay than luxury travelers for the first $\tilde{q}$ units of quality but not for additional quality/luxury.

In screening models with single crossing, see for example Bolton and Dewatripont (2005, ch. 2), quality is downward distorted for all but the highest type (compared to the symmetric information first best). This helps the principal (here: the airline) to extract rents. Or, as Dupuit (1849) explains for railway travel:

> It is not because of the few thousand francs that would have to be spend to put a roof over the third-class carriages or to upholster the third-class seats that some company or other has open carriages with wooden benches [...] What the company is trying to do is to prevent the passengers who can pay the second-class fare from traveling third-class.

Another standard result with single crossing is that only local incentive constraints are binding, i.e. with single crossing luxury travelers are indifferent between first class and business class and business travelers are indifferent between business and economy class. If single crossing is violated also non-local incentive constraints can bind: The solid lines in figure 4.1 depict indifference curves for the three types in a situation where the local downward constraints are binding.[64] But now–because of the violation of single crossing–the luxury traveler prefers economy to first class, i.e. his indifference curve through the first class offer lies above the economy class offer. Hence, a non-local incentive constraint is violated. The problem is that the price difference between economy and business class is determined by the huge utility difference of business travelers. The utility difference of luxury travelers between economy and business class quality is smaller than this price difference. To satisfy also the non-local incentive constraint, the contracts have to be

[63] Instead of collapsing all items in a one-dimensional quality index, one could alternatively look at a multidimensional model. See section 4.3 for the relation between the two approaches.

[64] The offers, i.e. quality and price, for each of the three classes are depicted as filled circles.

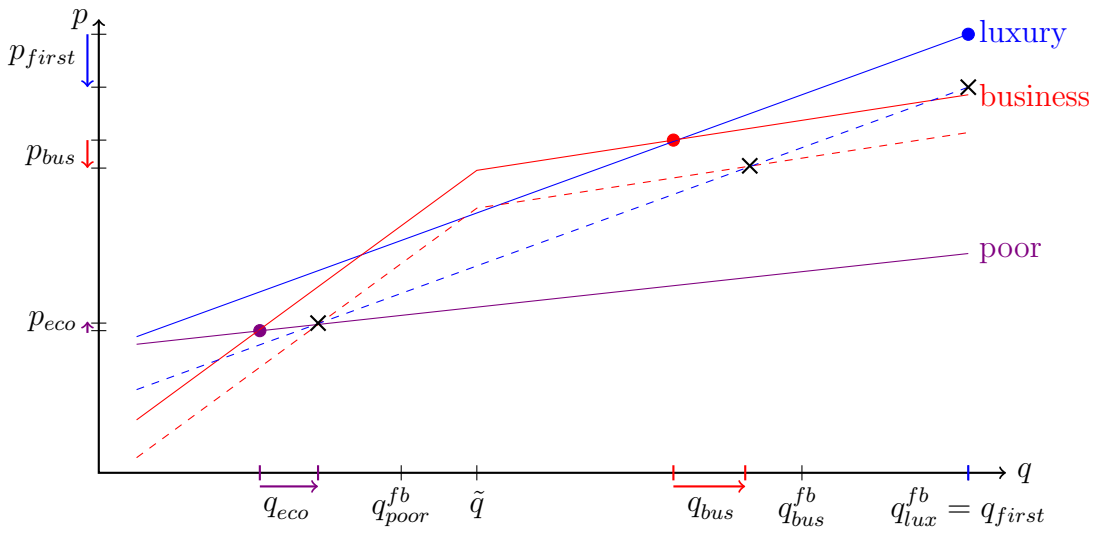


Figure 4.1: indifference curves: three type airline pricing example

changed to those indicated by "×" in fig 4.1. The dashed indifference curves in figure 4.1 go through these changed contracts.[65] Here, local and non-local incentive constraints bind. Essentially, two things happen when changing contracts: First, the price for first class travel has to decrease to prevent the luxury traveler from traveling economy class. As this also relaxes the incentive constraint between first and business class, the quality for business travelers can be brought closer to their first best quality. Second, the quality in the economy class increases. This helps to relax the non-local incentive constraint because it reduces the utility difference between economy and business class for business travelers more than for luxury travelers.

The main features of this example are (i) non-local incentive constraints can bind if single crossing is violated and (ii) binding non-local incentive constraints increase quality/reduce distortions. Both results will be generalized later on.

4.3. Literature

The standard screening model with single crossing is well known and explained in many textbooks, see for example Fudenberg and Tirole (1991) or Bolton and Dewatripont

[65] A completely solved numerical example corresponding to figure 4.1 can be found on `https://sites.google.com/site/christophschottmueller/research/webappendices`.

(2005). The literature on violations of single crossing in screening models remains relatively scarce.

Some insights have been gained for discrete type insurance models with perfect competition among principals. Several papers analyze settings where private information has two dimensions and can take either a high or a low value in each dimension, i.e there are 2×2 types. In Smart (2000), the two dimensions are risk and risk aversion while in Wambach (2000) they are wealth and risk. Netzer and Scheuer (2010) model an additional labor supply decision and the two dimensions are productivity and risk. All three papers share a pooling result, i.e. if single crossing is violated two of the four types can be pooled. Boone and Schottmüller (2011a) show that with imperfect competition among principals there can even be an order reversal: Types with higher risk can have more but also less insurance coverage if single crossing is violated.

My paper will analyze a model with a continuum of types and one principal. As I will illustrate in the next section, the main technical difficulty caused by a violation of single crossing are non-locally binding incentive constraints. In discrete type models one can take all incentive constraints explicitly into account. This is more difficult in a continuous type model since a continuum of constraints exists. Also, some additional qualitative results emerge from the continuous type model, e.g. distortion above as well as below first best and distortion at the top.

Araujo and Moreira (2010) characterize in a continuous type framework (inversely) U-shaped solutions in a setup where single crossing is not satisfied. In these solutions, some contracts are chosen by two types ("discrete pooling"). It turns out that in (inversely) U-shaped solutions non-local incentive constraints are only binding between types choosing the same contract from the menu. My paper complements their work by characterizing monotone solutions in the same model. The main technical difference is that non-local incentive constraints can bind between types choosing different contracts from the menu. Qualitatively, the solution in Araujo and Moreira (2010) features either a discontinuity or a bunching interval. Furthermore, there is a no distortion at the top result, i.e. the type with the highest first best decision[66] will be assigned this first best decision

[66]In Araujo and Moreira (2010), the function $s(\theta)$ is downward sloping and the analyzed solution is actually U-shaped (not *inversely* U-shaped as it would be when applied to my setting). In this setting, the undistorted top type is then the type with the *lowest* decision.

under the optimal contract. My paper shows that monotone solutions can be strictly monotone and continuous and therefore bunching and discontinuities are not a necessary implication of a violation of single crossing. Furthermore, I show that distortion a the top is possible in monotone solutions. The (inverse) U-shape solution and its critical condition (see section 4.5.1 and 4.8.1) is also applied in an insurance model (Araujo and Moreira, 2003), in signalling games (Araujo et al., 2007, 2011), non-linear monopoly pricing (Araujo et al., 2010) and auctions (Araujo et al., 2008).

Violations of single crossing are also related to the literature on multidimensional screening, see Armstrong (1996) and Rochet and Choné (1998) for seminal contributions, Rochet (2009) for a recent related paper and Rochet and Stole (2003) for a survey. As pointed out in the survey, "the problems arise not because of multiple dimensionality itself, but because of a commonly associated lack of exogenous type-ordering in multiple-dimensional environments." A violation of single crossing also conveys a lack of type-ordering. To make the relationship clear, think of a multidimensional, discrete type model. Clearly, one can reassign types to a one-dimensional parameter but this reassigned type will normally not satisfy single crossing. Consequently, an applied researcher will often have the choice between a multidimensional type model or a one-dimensional type model violating single crossing. My paper provides tools to make the latter way feasible.

The paper also relates to work relaxing the basic assumptions of the textbook model, e.g. Bolton and Dewatripont (2005, ch. 2). Jullien (2000) allows for type dependent participation constraints while Hellwig (2010) analyzes the case of irregular type distributions, i.e. distributions with mass points and zero densities. Hellwig (2010) shows that this leads to discontinuities as well as bunching in the optimal decision schedule. Contrary to my paper, there is no distortion above first best and no distortion at the top. In Jullien (2000), distortion can be above as well as below first best. The reason is that incentive constraints can bind upward as well as downward. If a participation constraint binds in the interior, it is relaxed by increasing the decision of lower types: This increases the slope of the rent function and leads to higher rents for the interior type. I show that binding non-local incentive constraints can also lead to distortion above first best although incentive constraints are only downward binding. The intuition is that if a type θ wants to misrepresent as a lower type $\hat{\theta}$, one can relax this non-local incentive constraint by increasing the decisions of types between $\hat{\theta}$ and θ: This will increase the

slope of the rent function and lead to higher rents for θ at his own contract. Hence, misrepresentation is less attractive.

There are also some papers in which violations of single crossing emerge but non-local incentive constraints are (assumed to be) non-binding due to the specific functional forms of the setup, see for example Calzolari (2004), Martimort and Stole (2009) or Hoffmann and Inderst (2011).

4.4. Model

There is a one-dimensional decision in a principal agent relationship which is denoted by $q \in \mathbb{R}_+$. Furthermore, there is a monetary transfer $t \in \mathbb{R}$. The agent's utility is $\pi = t - c(q, \theta)$ where $\theta \in \Theta \equiv [\underline{\theta}, \bar{\theta}] \subset \mathbb{R}$ is the type of the agent which is his private information. The function $c(q, \theta)$ is assumed to be three times continuously differentiable with $c_q > 0$, $c_\theta < 0$. The assumption $c_\theta < 0$ ensures that the participation constraint can only bind at the lowest type. Hence, any deviation from the standard solution will not be due to participation constraints binding in the interior, see Jullien (2000) for this, but to the violation of single crossing.

The principal's utility is $u(q, \theta) - t$ and is two times continuously differentiable with $u_q > 0$. The principal has the prior distribution $F(\theta)$ with continuous density $f(\theta) > 0$ for all $\theta \in [\underline{\theta}, \bar{\theta}]$. To simplify the exposition, I assume full participation, i.e. the surplus from trade is so high that it is not beneficial to exclude some types.[67]

For example, the principal could be the regulator of a natural monopolist and q could be the quality (or quantity) of service provided. The regulator might maximize expected consumer surplus which could be $q - p$ where p is the price paid. The natural monopolist would have cost function $c(q, \theta)$ and maximize profits. A higher type would correspond to a more efficient firm in the sense that its costs are lower than the costs of a lower type.

By the revelation principle, any general mechanism can also be implemented by a

[67] This is less restrictive than it might seem. By $c_\theta < 0$, only types at the low end could be excluded. If exclusion is optimal, the characterization in this paper applies to the set of not excluded types. Using the methods of this paper, one can calculate the solution for any given cutoff type and then maximize the principal's payoff over the cutoff type.

direct revelation mechanism in which the agent truthfully reports his type (Myerson, 1979). The task is to design a menu $q(\theta)$, implemented by transfers $t(\theta)$, which is individually rational (ir) and incentive compatible (ic) for the agent and maximizes the principal's objective under these two constraints.

Define $\pi(\theta)$ as the rents (in the regulation example: profits) a type θ gets under an implementable menu $(q(\theta), t(\theta))$. Faced with a menu $(q(\theta), t(\theta))$, a type θ agent will maximize $t(\hat{\theta}) - c(q(\hat{\theta}), \theta)$ over his type announcement $\hat{\theta}$. The envelope theorem and truthful revelation require $\pi_\theta(\theta) = -c_\theta(q(\theta), \theta)$.

Incentive compatibility of a decision $q(\theta)$ requires in general for any $\theta, \hat{\theta} \in \Theta$

$$\Phi(\theta, \hat{\theta}) \equiv \pi(\theta) - [\underbrace{\pi(\hat{\theta}) + c(q(\hat{\theta}), \hat{\theta})}_{t(\hat{\theta})} - c(q(\hat{\theta}), \theta)] \geq 0. \tag{IC}$$

Using the envelope condition above, $\Phi(\theta, \hat{\theta})$ can be rewritten as

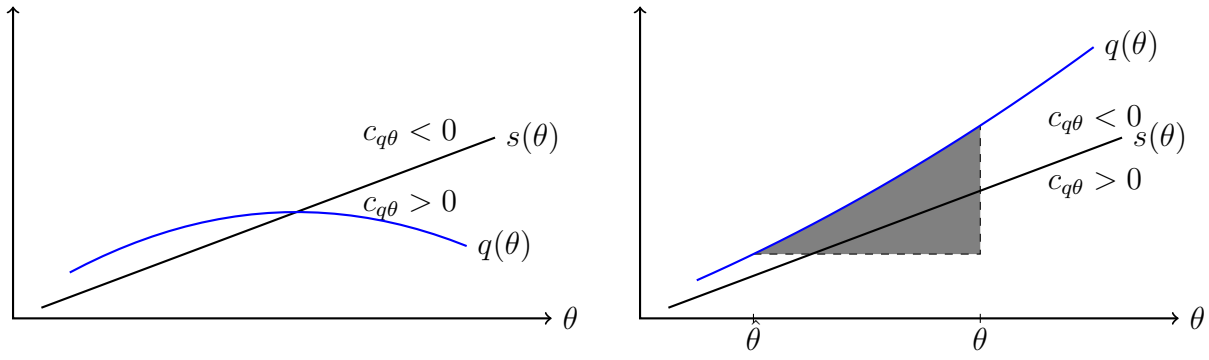
$$\Phi(\theta, \hat{\theta}) = \int_\theta^{\hat{\theta}} c_\theta(q(t), t) - c_\theta(q(\hat{\theta}), t)\, dt = -\int_\theta^{\hat{\theta}} \int_{q(t)}^{q(\hat{\theta})} c_{q\theta}(s, t)\, ds\, dt$$

where the second equality follows from integrating out the decision. Consequently, (IC) is equivalent to

$$-\int_\theta^{\hat{\theta}} \int_{q(t)}^{q(\hat{\theta})} c_{q\theta}(s, t)\, ds\, dt \geq 0. \tag{IC'}$$

Single crossing in this model is equivalent to $c_{q\theta}(q, \theta)$ not changing sign for any value of q and θ. But then incentive compatibility in (IC') boils down to a simple monotonicity condition on $q(\theta)$ (plus the envelope condition): If $c_{q\theta} < 0$, then inequality (IC') will hold whenever $q(\theta)$ is monotonically increasing. If however $c_{q\theta}$ can change sign, this is no longer true. It remains true that $q(\theta)$ has to be increasing (decreasing) at θ if $c_{q\theta}(q(\theta), \theta) < (>)0$. Otherwise, (IC') would be violated for types close enough to θ. But this no longer implies global incentive compatibility for two arbitrary types θ and $\hat{\theta}$.

This paper focusses on a one-time violation of single crossing also used by Araujo and Moreira (2010): It is assumed that $c_{q\theta}$ changes sign only once for a given q (or a given θ). More precisely, I assume $c_{q\theta\theta} > 0$ and $c_{qq\theta} < 0$. Hence, there exists a strictly increasing function $s(\theta)$ such that $c_{q\theta}(s(\theta), \theta) = 0$. Put differently, $s(\theta)$ gives for each type the decision level at which the cross derivative $c_{q\theta}$ is zero. The assumption on third derivatives are normally made to ensure concavity of the objective function and monotonicity of the decision, see for example section 7.3.2 in Fudenberg and Tirole



(a) inverse U-shape solution

(b) monotone solution

Figure 4.2: possible solution shapes

(1991). Here, however, they provide some structure on the way single crossing is violated. Note that in all examples of section 4.2.1 single crossing is violated only once.

As incentive compatibility requires $sgn(-c_{q\theta}(q(\theta),\theta)) = sgn(q_\theta(\theta))$, the optimal decision schedule cannot have arbitrary shapes. Araujo and Moreira (2010) analyze inverse U-shape decisions, see figure 4.2a. I will analyze monotone solutions in this paper, see figure 4.2b.[68]

Although $c_{q\theta} < 0$ at the decision $q(\theta)$ for all types, the violation of single crossing still plays a role in monotone solutions. It follows from (IC') that one can represent incentive compatibility as an integral over the shaded area in figure 4.2b: If the integral of $c_{q\theta}$ over this shaded area is negative, incentive compatibility is satisfied for θ and $\hat{\theta}$. Hence, the region where $c_{q\theta} > 0$ plays a role although the solution does not pass through this region.

The intuition is the following: Take two types θ and $\hat{\theta}$ with $\theta > \hat{\theta}$. Type $\hat{\theta}$ is assigned a transfer decision pair $(\hat{t}, \hat{q})$ and likewise θ has pair (t, q) with $q > \hat{q}$. When deciding whether he should misrepresent, type θ will compare the transfer difference $t - \hat{t}$ with the cost difference $c(q,\theta) - c(\hat{q},\theta)$. Note that the transfer difference $t - \hat{t}$ does not depend on type θ while the cost difference does. With single crossing, the cost difference is decreasing in type. If a type $\theta' \in (\hat{\theta}, \theta)$ with $q' \in (\hat{q}, q)$ is introduced, it follows that $c(q,\theta) - c(\hat{q},\theta) < c(q,\theta) - c(q',\theta) + c(q',\theta') - c(\hat{q},\theta')$. On the other hand, the equivalent

[68] I will only look at monotonically increasing solutions. It is easy to show that in solutions that are below $s(\theta)$ for all types (and therefore are decreasing) non-local incentive constraints do not bind. Hence, standard methods are sufficient for such problems.

expression for transfers holds with equality: $t - \hat{t} = t - t' + t' - \hat{t}$. Therefore, incentive compatibility between θ and $\hat{\theta}$ is implied by incentive compatibility between θ and θ' as well as between θ' and $\hat{\theta}$. Local incentive compatibility implies non-local incentive compatibility because single crossing implies that the cost difference is decreasing in type. Without the single crossing assumption, the cost difference $c(q,\theta) - c(\hat{q},\theta)$ is not necessarily decreasing in type and therefore local incentive constraints are not necessarily more demanding than non-local ones.

Before turning to the analysis of the solution, some definitions and one assumption is needed. I define the *first best* solution denoted by $q^{fb}(\theta)$ as the solution to

$$\max_{q(\theta)} u(q(\theta),\theta) - c(q(\theta),\theta)$$

which would be the optimal decision if the principal observed the agent's type. As a second reference point, it is useful to look at the *relaxed program*. This is the program taking only local incentive compatibility into account:

$$\begin{aligned} \max_{q(\theta)} \int_{\underline{\theta}}^{\bar{\theta}} &\{u(q(\theta),\theta) - c(q(\theta),\theta) - \pi(\theta)\} f(\theta)\, d\theta \qquad \text{(RP)} \\ s.t.:\ &\pi_\theta(\theta) = -c_\theta(q(\theta),\theta) \\ &q_\theta(\theta) c_{q\theta}(q(\theta),\theta) \leq 0 \\ &\pi(\theta) \geq 0 \end{aligned}$$

The first and second constraint are the local incentive compatibility constraints. More specifically, the first constraint is the envelope condition. It corresponds to a first order condition of the problem in which the agent maximizes his utility over his type announcement. The second constraint is the so called monotonicity constraint which corresponds to the second order condition of the same problem.[69] The third constraint is the participation constraint which will bind only for $\underline{\theta}$ by the assumption $c_\theta < 0$. I will call the solution of (RP) the *relaxed solution* and denote it by $q^r(\theta)$.

Since this paper focuses on the violation of single crossing in monotone solutions, the following assumption is made:

[69] For a brief proof of this (also for the case where single crossing is violated), see lemma 1 in Araujo and Moreira (2010).

Assumption 4.1. *The relaxed program is strictly concave in $q(\theta)$ and the relaxed solution is strictly monotonically increasing and strictly above $s(\theta)$, i.e. $q^r_\theta(\theta) > 0$ and $q^r(\theta) > s(\theta)$.*[70]

Put differently, I assume that the monotonicity constraint does not bind and the relaxed solution is fully characterized by the first order condition. It is easy to show that $u_{qq} \leq 0$ and $c_{qq} \geq 0$ are sufficient for concavity. For strict monotonicity and $q^r(\theta) > s(\theta)$, the following assumptions would be sufficient: $u_{q\theta} \geq 0$, $q^{fb}(\theta) > s(\theta)$ and the commonly made monotone hazard rate assumption, i.e. $f(\theta)/(1 - F(\theta))$ non-decreasing in θ. Note that the principal's utility is not influenced directly by the agent's type in most applications, i.e. even $u_{q\theta} = 0$ is often satisfied. The monotone hazard rate assumption is satisfied by the most common distributions as uniform, normal or exponential, see Bagnoli and Bergstrom (2005) for details. The assumption that $q^r(\theta)$– or equivalently $q^{fb}(\theta)$–is above $s(\theta)$ for all types sets the monotone case apart from the inverse U-shape case.

Under assumption 4.1, it is routine to verify that the relaxed solution is characterized by the first order condition

$$\{u_q(q(\theta), \theta) - c_q(q(\theta), \theta)\}f(\theta) + (1 - F(\theta))c_{q\theta}(q(\theta), \theta) = 0. \tag{4.3}$$

Since $q^r(\theta) > s(\theta)$, it follows that $c_{q\theta}(q^r(\theta), \theta) < 0$. Therefore, (4.3) implies that $q^r(\theta) \leq q^{fb}(\theta)$ where the inequality is strict for all types but $\bar{\theta}$. The fact that for the highest type $\bar{\theta}$ relaxed solution and first best coincide is the famous "no distortion at the top" result.

As already indicated, solutions can be monotone or inversely U-shaped (or even jumping over $s(\theta)$ discontinuously). It is therefore useful to have a sufficient condition under which the solution is monotone. To get such a sufficient condition, a technical condition has to be added to assumption 4.1. This technical assumption is sufficient to rule out solutions jumping discontinuously over $s(\theta)$. Given this, assumption 4.1 ensures that the solution is not inversely U-shaped but monotone.

To state this technical condition some "mirror images" have to be defined: Take a decision q below $s(\theta)$ and consider mirroring this decision in two ways: First, mirror it along $s(\theta)$: Define $q^s(q, \theta)$ by $\int_q^{q^s(q,\theta)} c_{q\theta}(x, \theta)\, dx = 0$. Second, mirror q along the relaxed

[70] Strict concavity of the relaxed program means that the partial derivative of the left hand side of equation(4.3) below with respect to q is negative.

solution q^r such that $\{u(q,\theta)-c(q,\theta)\}f(\theta)+(1-F(\theta))c_\theta(q(\theta),\theta)$ is the same for q and its mirror image $q^v(q,\theta)$. Since $c_\theta(q,\theta)$ and (RP) are concave in q, the two mirror images are well defined. Last define $q^f(\theta) < s(\theta)$ such that $q^s(q^f(\theta),\theta) = q^r(\theta)$, i.e. $q^f(\theta)$ is a kind of mirror image of the relaxed solution along $s(\theta)$.[71]

Appendix 4.8.3 shows that an optimal contract exists under the technical condition $q^v(q,\theta) \geq q^s(q,\theta)$ for all $q \in [0, q^f(\theta)]$ and all $\theta \in [\underline{\theta}, \bar{\theta}]$. The following proposition implies that this technical condition is–together with assumption 4.1– also sufficient for the monotonicity of the solution.

Proposition 4.1. *If $q^v(q,\theta) \geq q^s(q,\theta)$ for all $q \in [0, q^f(\theta)]$ and all $\theta \in [\underline{\theta}, \bar{\theta}]$, then any incentive compatible decision function $q(\theta)$ which imposes decisions below $s(\theta)$ for some type is dominated by the following changed decision*

$$q^c(\theta) = \begin{cases} q(\theta) & \text{if } q(\theta) \geq s(\theta) \\ q^s(q(\theta),\theta) & \text{if } q(\theta) < s(\theta) \end{cases}$$

combined with transfers such that $\pi^c_\theta = -c_\theta(q^c(\theta),\theta)$. This changed decision is above $s(\theta)$ and monotonically increasing. The changed decision is also incentive compatible.

Proof. see appendix

Put differently, the optimal decision has to be above $s(\theta)$ if the conditions of proposition 4.1 are met. To satisfy local incentive compatibility, a decision above $s(\theta)$ has to be monotonically increasing. The conditions of proposition 4.1 are therefore sufficient for monotonicity.

Note that the imposed condition is automatically satisfied for q close to $q^f(\theta)$ by assumption 4.1. Hence, the condition roughly states that $q^s(q,\theta)$ is not much steeper in q than $q^v(q,\theta)$. This holds, for example, true if $\{u(q,\theta)-c(q,\theta)\}f(\theta)+(1-F(\theta))c_\theta(q(\theta),\theta)$ and $c_\theta(\cdot)$ are both symmetric in q as then $q^v_q(q,\theta) = q^s_q(q,\theta) = -1$.

To illustrate, take example 1 from section 4.2.1 and assume that $u(q,\theta) = Sq$ with $S > 0$. In this case, $q^v(q,\theta)$ is defined by the equation

$$\left(S - \theta + \frac{1-F(\theta)}{f(\theta)}\right)(q - q^v) - \left(\frac{1}{\theta} + \frac{1-F(\theta)}{\theta^2 f(\theta)}\right)(q^2 - q^{v2}) = 0.$$

Straightforward calculation shows that this is solved by $q^v(q,\theta) = 2q^r(\theta) - q$. Hence, $q^v_q(q,\theta) = -1$ regardless of the type distribution and the conditions of the proposition are satisfied whenever assumption 4.1 holds.

[71] If no $q^f(\theta) \geq 0$ exists, take $q^f(\theta) = 0$.

Given that the condition $q^v(q,\theta) \geq q^s(q,\theta)$ is sufficient but not necessary for a monotone solution, this condition will not be used in the remainder of the paper where monotone solutions are characterized.

4.5. Optimal contract

4.5.1. Necessary conditions

This subsection presents necessary conditions which have to be met whenever a non-local incentive constraint is binding. Since these conditions are only a slight generalization of those presented in Araujo and Moreira (2010), the presentation will be brief and more intuitive than formal.

Take an optimal decision schedule $q(\theta)$ and let transfers be determined by local incentive compatibility, i.e. such that $\pi_\theta(\theta) = -c_\theta(q(\theta),\theta)$ and $\pi(\underline{\theta}) = 0$. Furthermore, suppose that IC is binding for two types θ and $\hat{\theta}$, i.e. $\Phi(\theta,\hat{\theta}) = 0$. By incentive compatibility, $\Phi(\cdot)$ has to be non-negative for all types. Therefore, $(\theta,\hat{\theta}) \in argmin_{(s,t)}\Phi(s,t)$ as $\Phi(\theta,\hat{\theta}) = 0$.

Given that $\pi(\cdot)$ and $c(\cdot)$ are differentiable, the first order condition with respect to θ has to hold:[72]

$$\frac{\partial\Phi(\theta,\hat{\theta})}{\partial\theta} = -c_\theta(q(\theta),\theta) + c_\theta(q(\hat{\theta}),\theta) \leq 0 \qquad \text{with "=" if } \theta < \bar{\theta} \tag{C1}$$

In the same way the first order condition for $\hat{\theta}$ is derived:[73]

$$\frac{\partial\Phi(\theta,\hat{\theta})}{\partial\hat{\theta}} = q_\theta(\hat{\theta})\left(-c_q(q(\hat{\theta}),\hat{\theta}) + c_q(q(\hat{\theta}),\theta)\right) \geq 0 \qquad \text{with "=" if } \hat{\theta} > \underline{\theta} \tag{C2}$$

Hence, $\hat{\theta}$ is either bunched or marginal costs of θ and $\hat{\theta}$ are equal at $q(\hat{\theta})$.

The interpretation of these two conditions is the following: Recall that $\pi_\theta(\theta) = -c_\theta(q(\theta),\theta)$ while $c_\theta(q(\hat{\theta}),\theta)$ is how profits of misrepresenting as $\hat{\theta}$ change in the misrepresenting type θ. Then condition (C1) says that profits $\pi(\theta)$ should change in type in the

[72] It turns out that non-local incentive compatibility constraints are only downward binding, see lemma 4.1. For this reason as well as notational convenience, I ignore the possibilities $\Phi(\theta,\bar{\theta}) = 0$ and $\Phi(\underline{\theta},\hat{\theta}) = 0$ already here.

[73] Differentiability of $q(\cdot)$ at $\hat{\theta}$ is not essential for this condition; see theorem 1A in Araujo and Moreira (2010) and the graphical interpretation below.

same way as misrepresentation-profits change in type. For a graphical interpretation, it is worthwhile to rewrite (C1) as

$$\int_{q(\hat{\theta})}^{q(\theta)} c_{q\theta}(q,\theta)\,dq = 0 \tag{C1'}$$

which means that the right hand side boundary of the shaded area in figure 4.2b is zero when weighted with $c_{q\theta}$. If the integral above was positive and $\Phi(\theta,\hat{\theta}) = 0$, then incentive compatibility would be violated for $\theta+\varepsilon$ and $\hat{\theta}$ as $\Phi(\theta+\varepsilon,\hat{\theta}) \approx \Phi(\theta,\hat{\theta}) - \varepsilon \int_{q(\hat{\theta})}^{q(\theta)} c_{q\theta}(q,\theta)\,dq$, i.e. the "shaded area" for $\theta+\varepsilon$ would be the same plus some area having the "wrong" sign.

If the integral above is negative, the same applies accordingly for $\theta-\varepsilon$, i.e. $\Phi(\theta-\varepsilon,\hat{\theta}) \approx \Phi(\theta,\hat{\theta}) + \varepsilon \int_{q(\hat{\theta})}^{q(\theta)} c_{q\theta}(q,\theta)\,dq$.

The second condition simply says that either $\hat{\theta}$ is bunched with other types or also the weighted lower boundary of the shaded area in figure 4.2b is zero, i.e.

$$\int_{\hat{\theta}}^{\theta} c_{q\theta}(q(\hat{\theta}),t)\,dt = 0. \tag{C2'}$$

Again, figure 4.2b illustrates the idea. If the integral was positive, incentive compatibility would be violated between θ and $\hat{\theta}-\varepsilon$ as $\Phi(\theta,\hat{\theta}-\varepsilon) \approx \Phi(\theta,\hat{\theta}) - \varepsilon q_\theta(\hat{\theta}) \int_{\hat{\theta}}^{\theta} c_{q\theta}(q(\hat{\theta}),t)\,dt$.

The graphical interpretation also allows to quickly generalize these conditions at points of discontinuity and bunching. This situation is depicted in figure 4.3. Assume $\Phi(\theta,\hat{\theta}_i) = 0$ for $i = 1,2$. To keep incentive compatibility for types close to θ, $\hat{\theta}_1$ and $\hat{\theta}_2$ the following conditions have to hold:[74]

- $\int_{q(\hat{\theta}_i)}^{q^-(\theta)} c_{q\theta}(q,\theta)\,dq \geq 0$ as otherwise $\Phi(\theta-\varepsilon,\hat{\theta}_i) < 0$
- $\int_{q(\hat{\theta}_i)}^{q^+(\theta)} c_{q\theta}(q,\theta)\,dq \leq 0$ as otherwise $\Phi(\theta+\varepsilon,\hat{\theta}_i) < 0$
- $\int_{\hat{\theta}_1}^{\theta} c_{q\theta}(q(\hat{\theta}),t)\,dt \leq 0$ as otherwise $\Phi(\theta,\hat{\theta}_1-\varepsilon) < 0$
- $\int_{\hat{\theta}_2}^{\theta} c_{q\theta}(q(\hat{\theta}),t)\,dt \geq 0$ as otherwise $\Phi(\theta,\hat{\theta}_1+\varepsilon) < 0$

Given (C1) and (C2), one can use variational calculus to derive a third necessary condition for types at which the incentive constraint binds. While (C1) and (C2) are purely driven by incentive compatibility, this third condition will be derived from the principal's optimization. The idea is to to have a variation of the optimal decision

[74] I use the superscript "–" ("+") to indicate limits from below (above).

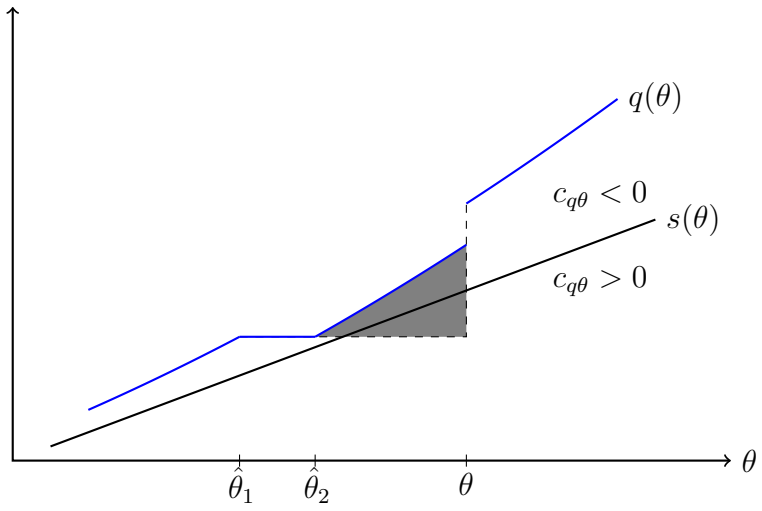


Figure 4.3: necessary conditions at discontinuity

around θ and $\hat{\theta}$ such that the two necessary conditions (C1) and (C2) are still satisfied. The method differs only slightly from the one used in Araujo and Moreira (2010) for discretely pooled types and therefore the steps are relegated to appendix 4.8.1. The following variational condition results:

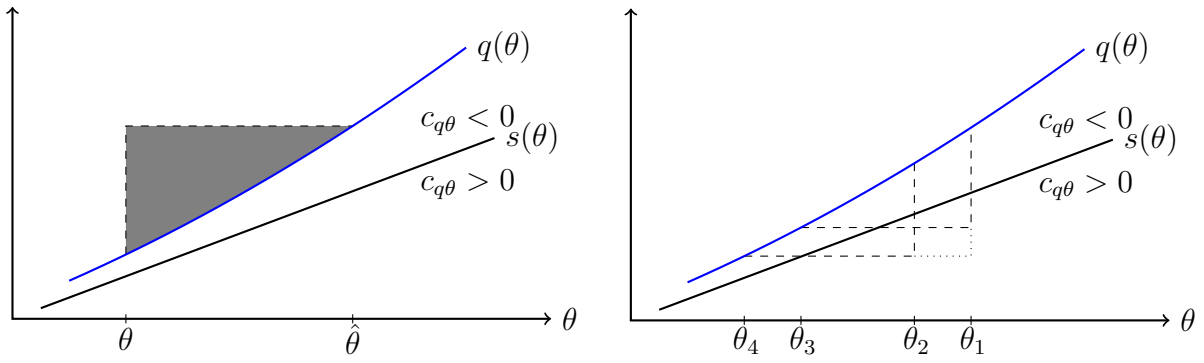
$$\frac{[u_q(q(\theta),\theta) - c_q(q(\theta),\theta)]f(\theta)}{c_{q\theta}(q(\theta),\theta)} + 1 - F(\theta) = \frac{[u_q(q(\hat{\theta}),\hat{\theta}) - c_q(q(\hat{\theta}),\hat{\theta})]f(\hat{\theta})}{c_{q\theta}(q(\hat{\theta}),\hat{\theta})} + 1 - F(\hat{\theta}) \quad \text{(C3)}$$

Section 4.5.2 will give a shadow value interpretation to the terms on both sides of (C3) and thereby provide some intuition for this condition.

4.5.2. Monotone solution

The remainder of the paper deals with the characterization of monotone solutions. As pointed out before, the main difficulties are non-locally binding incentive constraints. The following two lemmata show that only a certain subset of non-local incentive constraints can be binding. Lemma 4.1 implies that incentive constraints cannot be upward binding in monotone solutions. Put differently, no type will be indifferent between the contract designated for him and the contract of a higher type. The only possible way a non-local incentive constraint can be binding is downward, i.e. a type might be indifferent between his contract and the contract of a lower type.

Lemma 4.1. *If $q(\theta) \geq s(\theta)$ and $q(\theta)$ is locally incentive compatible, then no type wants to (non-locally) misrepresent upwards.*



(a) no upwards binding

(b) no overlap

Figure 4.4: non-binding constraints

Proof. Recall that local incentive compatibility requires monotonicity of $q(\theta)$, i.e. $q(\theta)$ has to be monotonically increasing as $q(\theta) \geq s(\theta)$. Now take $\hat{\theta} > \theta$. Incentive compatibility requires

$$\Phi(\theta, \hat{\theta}) \equiv \pi(\theta) - \pi(\hat{\theta}) - c(q(\hat{\theta}), \hat{\theta}) + c(q(\hat{\theta}), \theta) \geq 0. \tag{4.4}$$

Because of local incentive compatibility, this can be rewritten as

$$\int_{\theta}^{\hat{\theta}} c_{\theta}(q(t), t) - c_{\theta}(q(\hat{\theta}), t)\, dt = -\int_{\theta}^{\hat{\theta}} \int_{q(t)}^{q(\hat{\theta})} c_{q\theta}(s, t)\, ds\, dt \geq 0.$$

But the last inequality holds automatically since $q(\theta) \geq s(\theta)$ and $q_{\theta}(\theta) \geq 0$. This implies that the integrand is non-positive for all (s, t) in question. Figure 4.4a gives a graphical representation of this fact. *Q.E.D.*

The intuition for lemma 4.1 is the same as in models with single crossing. A higher decision increases the costs for higher types less than for lower types. For a low type, this holds true for all decisions above his own. Local incentive compatibility induces transfer differences making higher types indifferent between their decision and a marginally higher decision. A lower type will face the same transfer differences but higher cost differences when opting for a higher decision. Therefore, local incentive compatibility of higher types implies that low types do not want to misrepresent upwards non-locally.

The following lemma puts more structure on the ways incentive compatibility constraints can bind. It states that binding non-local incentive constraints cannot overlap, i.e. if a non-local incentive constraint binds between θ and $\hat{\theta}$, then no other incentive constraint can bind between a type in the interval $[\hat{\theta}, \theta]$ and a type outside this interval.

I will use the following phrase to describe binding non-local incentive constraints: A non-local incentive constraint binds *from* θ *to* $\hat{\theta}$ if $\Phi(\theta, \hat{\theta}) = 0$.

Lemma 4.2. *Assume the solution is monotone and assume* $q(\hat{\theta}) < q(\theta)$. *If the non-local incentive constraint binds from* θ *to* $\hat{\theta}$, *it cannot bind from any* $\theta' \in [\hat{\theta}, \theta)$ *to any* $\hat{\theta}' \notin [\hat{\theta}, \theta)$. *Neither can it bind for any* $\hat{\theta}'' \in (\hat{\theta}, \theta]$ *and* $\theta'' \notin (\hat{\theta}, \theta)$.

Proof. The proof is by contradiction. Suppose, contrary to the lemma, there are types $\theta_1 > \theta_2 \geq \theta_3 > \theta_4$ with $\Phi(\theta_1, \theta_3) = 0$ and $\Phi(\theta_2, \theta_4) = 0$. Then the incentive constraint between θ_1 and θ_4 will be violated, i.e. $\Phi(\theta_1, \theta_4) < 0$:

$$\begin{aligned}
\Phi(\theta_1, \theta_4) &= -\int_{\theta_4}^{\theta_1} \int_{q(\theta_4)}^{q(t)} c_{q\theta}(s,t)\, ds\, dt \\
&= -\int_{\theta_4}^{\theta_2} \int_{q(\theta_4)}^{q(t)} c_{q\theta}(s,t)\, ds\, dt - \int_{\theta_2}^{\theta_1} \int_{q(\theta_4)}^{q(\theta_3)} c_{q\theta}(s,t)\, ds\, dt - \int_{\theta_2}^{\theta_1} \int_{q(\theta_3)}^{q(t)} c_{q\theta}(s,t)\, ds\, dt \\
&= -\int_{\theta_4}^{\theta_2} \int_{q(\theta_4)}^{q(t)} c_{q\theta}(s,t)\, ds\, dt - \int_{\theta_2}^{\theta_1} \int_{q(\theta_4)}^{q(\theta_3)} c_{q\theta}(s,t)\, ds\, dt \\
&\quad + \int_{\theta_3}^{\theta_2} \int_{q(\theta_3)}^{q(t)} c_{q\theta}(s,t)\, ds\, dt - \int_{\theta_3}^{\theta_1} \int_{q(\theta_3)}^{q(t)} c_{q\theta}(s,t)\, ds\, dt \\
&= -\Phi(\theta_2, \theta_3) - \int_{\theta_2}^{\theta_1} \int_{q(\theta_4)}^{q(\theta_3)} c_{q\theta}(s,t)\, ds\, dt < 0
\end{aligned}$$

The first and second equality are simple splitting up the integral steps (and can readily be seen in figure 4.4b), the third uses the fact that $\Phi(\theta_1, \theta_3) = \Phi(\theta_2, \theta_4) = 0$ and the last inequality follows from the incentive compatibility between θ_2 and θ_3 as well as the following idea: By the binding constraint between θ_2 and θ_4 and the fact that θ_2 is interior, $\int_{q(\theta_4)}^{q^-(\theta_2)} c_{q\theta}(s, \theta_2)\, ds \geq 0$ holds by C1 (with equality if $q(\theta)$ is continuous at θ_2). By the monotonicity of $q(\cdot)$, $q(\theta_3) \leq q^-(\theta_2)$ and therefore $\int_{q(\theta_4)}^{q(\theta_3)} c_{q\theta}(s, \theta_2)\, ds \geq 0$ (see figure 4.4b). The inequality above follows then from $c_{q\theta\theta} \geq 0$. *Q.E.D.*

As a special case, i.e. with $\hat{\theta} = \theta'$, the preceding lemma includes the following: If θ is indifferent between his and $\hat{\theta}$'s contract, i.e. $\Phi(\theta, \hat{\theta}) = 0$, then no other type θ' is indifferent between his contract and θ's contract, i.e. $\Phi(\theta', \theta) > 0$ for all $\theta' \in \Theta \setminus \theta$. Put differently, incentive compatibility can bind non-locally from a type or to a type but not both. Figure 4.5 summarizes the two previous lemmata by showing how non-local incentive compatibility constraints can bind in a monotone solution.

One of the contributions of this paper is that a violation of single crossing can affect the solution without leading to irregularities, i.e. discontinuities or bunching. The fol-

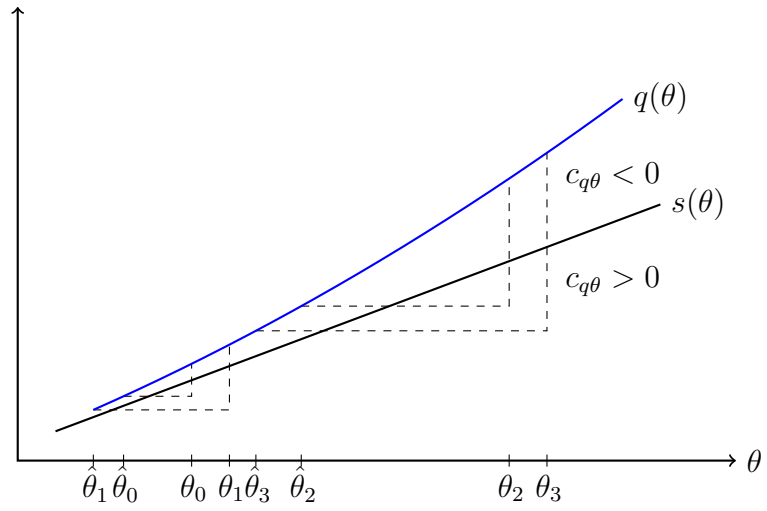


Figure 4.5: how incentive constraints can bind

lowing lemma shows that some irregularities can be ruled out on the grounds of incentive compatibility alone.

Lemma 4.3. *Assume a non local incentive constraint binds from θ to $\hat{\theta}$, i.e. $\Phi(\theta, \hat{\theta}) = 0$. The decision is continuous at $\hat{\theta}$ if $\hat{\theta}$ is not the boundary type of a bunching interval. Furthermore, θ cannot be bunched if the decision is continuous at θ and $\theta < \bar{\theta}$.*

Proof. see appendix

After these technical results, it is possible to obtain a qualitative result of practical importance. If the solution is monotone, non-local incentive compatibility might require "distortions" that are unusual: *With single crossing,* local incentive constraints are downward binding. This explains why the relaxed solution is below the first best decision. With single crossing, a high type has lower marginal costs than a low type. By distorting the low type's decision downward, the cost advantage of the high type is reduced, i.e. the low type's decision becomes less attractive. Consequently, the rent paid to the high type can be lower without inducing misrepresentation. *Without single crossing,* it is no longer clear that a high type has lower marginal costs than a low type at the low type's decision. Figure 4.2b, for example, illustrates that $\int_{\hat{\theta}}^{\theta} c_{q\theta}(q(\hat{\theta}), t)\, dt = c_q(q(\hat{\theta}), \theta) - c_q(q(\hat{\theta}), \hat{\theta})$ could be positive. Therefore, making the low type's contract unattractive might require increasing the low type's decision. Informational distortion from local and non-local incentive constraints will then go in opposite directions. In short, binding non-local incentive constraints reduce the usual downward distortions.

Proposition 4.2. *If the optimal decision is monotone, it will be above the relaxed solution, i.e. if $q(\theta)$ is monotonically increasing, then $q(\theta) \geq q^r(\theta)$ for all θ.*

Proof. see appendix

The previous proposition highlights how violations of non-local ic are dealt with under monotone solutions. This can also be illustrated with figure 4.2b. Incentive compatibility is violated if the grey area weighted by $c_{q\theta}$ is positive. To satisfy incentive compatibility one can raise q for all types between $\hat{\theta}$ and θ. The additional grey area features $c_{q\theta} < 0$ and therefore the incentive problem is mitigated.

One noteworthy point is that the incentive constraint is mainly relaxed by increasing q for types at which the incentive constraint is non-binding; i.e. if ic is binding from θ' to $\hat{\theta}'$, it is less $q(\theta')$ and $q(\hat{\theta}')$ that have to be increased but q for the types between $\hat{\theta}'$ and θ'. To see the intuition, recall that $\pi_\theta(\theta) = -c_\theta(q(\theta), \theta)$ and that $c_{q\theta}(q(\theta), \theta) < 0$. Therefore, increasing q will raise the slope of the rent function $\pi(\theta)$. Increasing q for types in $(\hat{\theta}', \theta')$ will therefore increase the rent of θ' at his assigned menu point. Obviously, the non-local incentive constraint is relaxed.

The last paragraph illustrates that non-local incentive constraints are potentially difficult to handle: The decision of a type is not only influenced by the incentive constraints binding for him but also by binding incentive constraints of other types. The following theorem structures this intuition and characterizes the solution.

Theorem 4.1. *A monotone solution is characterized by the equation*

$$[u_q(q(\theta), \theta) - c_q(q(\theta), \theta)]f(\theta) + (1 - F(\theta))c_{q\theta}(q(\theta), \theta) = \eta(\theta)c_{q\theta}(q(\theta), \theta) \tag{4.5}$$

where $\eta(\theta)$ is a non-negative function with the following properties:

- *$\eta(\theta)$ is constant on each interval of types for which non-local incentive constraints are not binding and the decision is strictly increasing.*
- *$\eta(\theta)$ is non-decreasing at types $\hat{\theta}$ to which non-local incentive constraints are binding whenever $\hat{\theta}$ is not bunched.*
- *$\eta(\theta)$ is non-increasing at types from which non-local incentive constraints are binding.*
- *$\eta(\bar{\theta})$ is zero if no non-local incentive constraint is binding from $\bar{\theta}$.*

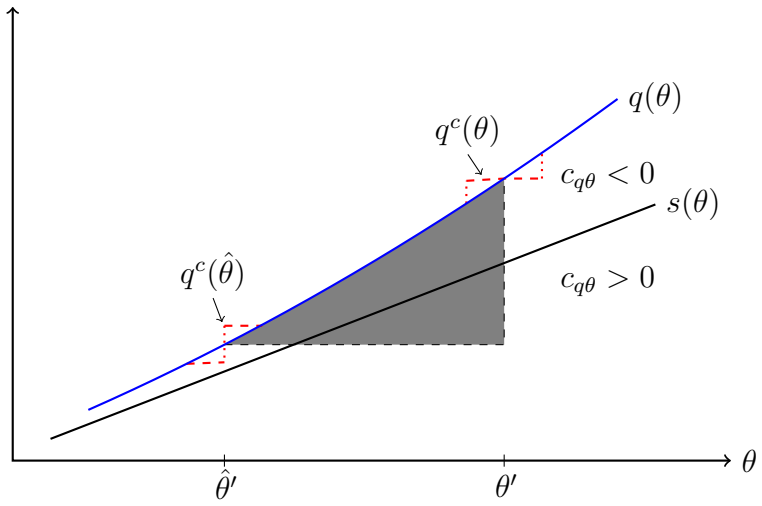


Figure 4.6: feasible changes

- *$\eta(\underline{\theta})$ is zero if no non-local incentive constraint is binding to $\underline{\theta}$.*

Proof. see appendix

Before giving an intuitive interpretation to $\eta(\theta)$, let me briefly sketch the idea behind the proof of the theorem. Given the solution $q(\theta)$, one can simply define $\eta(\theta)$ by (4.5). The properties of $\eta(\theta)$ are derived by showing that $q(\theta)$ could be changed in a way that (i) is incentive compatible and (ii) increases the principal's payoff if these properties were not satisfied. Figure 4.6 shows feasible changes when a non-local incentive constraint is binding from θ' to $\hat{\theta}'$. Increasing the decision for types slightly below θ' will relax (or not affect) binding non-local incentive constraints. Since this change relaxes the incentive constraints from types above θ' to types below θ', it is then feasible to assign types slightly above θ' a lower decision, see figure 4.6. Note that lemma 4.2 is essential for feasibility as it assures that no non-local incentive constraint is binding *to* types slightly above θ'. It can then be shown that such a feasible change would increase the principal's payoff if $\eta(\theta)$ was increasing at θ'. At $\hat{\theta}$, a different change in the decision is feasible, see figure 4.6, which can be used to show that $\eta(\theta)$ cannot be decreasing at $\hat{\theta}$. At types where non-local incentive constraints are lax, both kind of changes are feasible and consequently $\eta(\theta)$ has to be constant.

The properties of $\eta(\theta)$ have an intuitive interpretation. The left hand side of (4.5) is well known from models with single crossing: Increasing $q(\theta)$ affects the surplus from type θ but also the rents of all types above θ. But marginally increasing $q(\theta)$ will also relax

all non-local incentive constraints binding from types $\theta' > \theta$ to types $\hat{\theta}' < \theta$, see figure 4.2b. As these incentive constraints can be expressed as integrals over $c_{q\theta}$ (see equation (IC')), the "amount" by which those non-local incentive constraints are relaxed is given by $c_{q\theta}(q(\theta), \theta)$ which can be found on the right hand side of (4.5). Consequently, $\eta(\theta)$ could be interpreted as the shadow value of all non-local incentive constraints binding from types $\theta' > \theta$ to types $\hat{\theta}' < \theta$. These binding constraints are the same for all types in an interval of types for which non-local incentive constraints are lax, see figure 4.5. This explains the first property of $\eta(\theta)$.

The other properties can also be explained by the shadow value interpretation of $\eta(\theta)$. If a non-local incentive constraint is binding to a type $\hat{\theta}$, then there are more non-local incentive constraints binding "over" $\hat{\theta} + \varepsilon$ than "over" $\hat{\theta} - \varepsilon$.[75] Consequently, the shadow value of non-local incentive constraints binding over a type has to be higher for $\hat{\theta}+\varepsilon$ than for $\hat{\theta}-\varepsilon$. Put differently, increasing $q(\hat{\theta}+\varepsilon)$ relaxes more non-local incentive constraints than increasing $q(\hat{\theta} - \varepsilon)$.

Also the last two properties are straightforward: Increasing the decision of the boundary types does not affect non-local incentive constraints of other types.

Furthermore, the interpretation as shadow value provides some intuition for the necessary condition (C3) which basically says that $\eta(\theta) = \eta(\hat{\theta})$ when a non-local incentive constraint is binding from θ to $\hat{\theta}$. This makes sense in light of lemma 4.2. Because there is no overlap in binding incentive constraints, the non-local incentive constraints binding over θ are the same as the ones binding over $\hat{\theta}$. Consequently, the shadow value of relaxing those constraints is the same for the two types.

Theorem 4.1 establishes what happens at types where non-local incentive constraints are binding (or lax). Here I want to argue that non-local incentive constraints are typically binding from and to intervals of types. Put differently, there are intervals $[\theta_0, \theta_1]$ and $[\hat{\theta}_1, \hat{\theta}_0]$ such that a non-local incentive constraint is binding from each $\theta' \in [\theta_0, \theta_1]$ to some $\hat{\theta}' \in [\hat{\theta}_1, \hat{\theta}_0]$. From theorem 4.1, it follows that $\eta(\theta') = \eta(\hat{\theta}')$ and $\eta(\theta)$ is increasing (decreasing) on $[\hat{\theta}_1, \hat{\theta}_0]$ (on $[\theta_0, \theta_1]$). The intuition for this structure is the following: Take types θ' and $\hat{\theta}'$ such that a non-local incentive constraint between θ' and $\hat{\theta}'$ is violated under the relaxed solution. Proposition 4.2 indicates that the decision of the types between $\hat{\theta}$ and θ' is increased to establish incentive compatibility. The usual optimization

[75] With binding "over" θ I mean binding from a type $\theta' > \theta$ to a type $\hat{\theta} < \theta$.

intuition suggests that it should be optimal to increase the decision for all those types by "the same amount."[76] However, this is not possible because of incentive compatibility constraints: Clearly, the decision of types $\theta' - \varepsilon$ cannot be increased discretely because of the monotonicity constraint at θ'. Lemma 4.3 establishes that the monotonicity constraint cannot even be binding for θ' as then the non-local constraint from $\theta' - \varepsilon$ to $\hat{\theta}'$ would be violated. Lemma 4.3 also makes clear that the decision should not jump at $\hat{\theta}'$ as otherwise the non-local constraint from θ' to $\hat{\theta}' + \varepsilon$ would be violated. One could now conjecture that non-local incentive constraints are binding from θ' not only to $\hat{\theta}'$ but also to slightly higher types and–with the same logic–from types slightly below θ' to $\hat{\theta}'$. However, it is not difficult to show that the incentive constraint between $\theta' - \varepsilon$ and $\hat{\theta}' + \varepsilon$ would be violated in this case. Consequently, one is left with the interval structure described above where non-local incentive constraints are binding from types slightly below θ' to types slightly above $\hat{\theta}'$.

The following lemma takes another perspective on the structure by establishing that non-local incentive constraints cannot bind at a finite number of interior types. With the additional properties established in the lemma, one should indeed expect the set of types where non-local incentive constraints bind to contain an interval.[77]

Lemma 4.4. *If the optimal solution is monotone and the relaxed solution is not implementable*[78]*, non-local incentive constraints cannot bind only from a finite number of interior types to a finite number of interior types. The set of types from (to) which non-local incentive constraints bind cannot consist of isolated interior types.*[79]

The solution can be chosen such that (i) the set of types from which non-local incentive constraints are binding is closed and (ii) the set of types to which non-local incentive constraints are binding is closed.

Proof. see appendix

[76] Theorem 4.1 confirms this intuition by establishing that $\eta(\theta)$ is constant at types where non-local incentive constraints are lax.

[77] Strictly speaking, the lemma leaves the option that non-local incentive constraints are binding at a Cantor set of interior types. As the following results do not depend on this artificial looking case, I will ignore this possibility and speak of intervals in the remainder of the paper.

[78] The relaxed decision is said to be *not implementable* if it violates non-local incentive constraints.

[79] Isolated means here that for each type θ from (to) which a non-local incentive constraint binds, there exists a neighborhood of θ in which non-local incentive constraints are lax for all types but θ.

Some of the properties of $\eta(\theta)$ in theorem 4.1 hold only at types where the decision is strictly increasing. The reason is that, the way (4.5) is written, $\eta(\theta)$ captures not only the effect of non-local incentive constraints but also the effect of the monotonicity constraint. If one wants to avoid this cluttering of effects, it is straightforward to introduce a monotonicity parameter $\nu(\theta)$ which captures the effect of the monotonicity constraint. In this case it is easy to see that the properties of $\eta(\theta)$ described in theorem 4.1 extend also to bunched types. Instead of (4.5) the solution would then be characterized by

$$\nu_\theta(\theta) = (u_q(q(\theta),\theta) - c_q(q(\theta),\theta))f(\theta) + (1 - F(\theta) - \eta(\theta))c_{q\theta}(q(\theta),\theta)$$

where $\nu(\theta)q_\theta(\theta) = 0$ for all $\theta \in \Theta$, i.e. $\nu(\theta)$ corresponds to the Lagrange parameter of the monotonicity constraint. If the start and ending type of a bunching interval are denoted by θ_s^b and θ_e^b, then obviously $\int_{\theta_s^b}^{\theta_e^b} \nu_\theta(\theta)\, d\theta = 0$. As described in the existing literature on ironing, see Guesnerie and Laffont (1984) or the exposition in Fudenberg and Tirole (1991, ch. 7), the bunching interval is characterized by this last condition and the endpoint conditions $\nu(\theta_s^b) = \nu(\theta_e^b) = 0$. The following lemma formalizes the discussion of the last paragraph.

Lemma 4.5. *If types in the interval (θ_s^b, θ_e^b) are bunched in the optimal solution, then there exists a function $\eta(\theta)$ which satisfies the properties of theorem 4.1 also for bunched types. In particular, $\eta(\theta)$ is non-decreasing on (θ_s^b, θ_e^b) and constant if no non-local incentive constraint binds to the bunched types. Furthermore, $\eta(\theta)$ satisfies (i) $\eta(\theta) = \eta(\hat{\theta})$ if $\Phi(\theta,\hat{\theta}) = 0$ and (C1') as well as (C2') hold, (ii) $\int_{\theta_s^b}^{\theta_e^b} \nu_\theta(\theta)\, d\theta = 0$ with $\nu_\theta(\theta)$ defined as above.*

Proof. see appendix

4.5.3. Continuous solutions

This subsection has two goals: First, to provide sufficient conditions under which a monotone solution is continuous and, second, to introduce an algorithm for determining such a continuous solution.

The first sufficient condition for continuity is loosely based on the idea of having a one-to-one relationship between η and q for a given type θ; i.e. the idea that for a given type θ and $\eta(\theta) > 0$, equation (4.5) yields a unique solution for q. The condition

in the proposition ensures this and also ascertains that this relationship is monotonic, i.e. a higher $\eta(\theta)$ results in a higher q. Furthermore, the same condition ensures that the principal's objective is concave when the optimization problem is written as an optimization over the rent profile $\pi(\theta)$ instead as an optimization over $q(\theta)$, see the proof of proposition 4.6.

Proposition 4.3. *Assume that the first best is concave, i.e.* $u_{qq} - c_{qq} \leq 0$. *A monotone solution is continuous if*

$$\frac{u_{qq}(q,\theta) - c_{qq}(q,\theta)}{c_{qq\theta}(q,\theta)} > \frac{u_q(q,\theta) - c_q(q,\theta)}{c_{q\theta}(q,\theta)} \tag{CVR}$$

holds for all types and all $q \geq q^{fb}(\theta)$.[80]

Proof. see appendix

Hence, if the social objective $u(q,\theta) - c(q,\theta)$ is concave enough or if the cross derivative $c_{q\theta}(q,\theta)$ is in absolute value large enough (at the first best decision), the optimal decision will be continuous. Take for example the cost function in example 1 in section 4.4 and assume that $u(q,\theta) = \beta q$. It turns out that (CVR) is equivalent to the condition for $q^{fb}(\theta) > s(\theta)$, i.e. $\beta > 2\bar{\theta}$.[81]

The following proposition gives an alternative condition under which the optimal solution is below the first best decision. Having a solution below first best turns out to be sufficient for continuity and strict monotonicity of the solution (under a standard monotone hazard rate assumption). This is in itself remarkable. As the relaxed solution is below first best, one should expect the solution to be below first best whenever non-local incentive constraints are not violated "too much" by the relaxed solution. Hence, there is a broad class of problems in which the solution will be strictly monotone and continuous. Furthermore, the proof of the following proposition shows that the property holds also locally. That is, if the decision is below first best on some interval (θ_1, θ_2), then the decision will be strictly monotone and continuous on (θ_1, θ_2).

Before stating the proposition some additional notation is needed. Define $q^m(\theta)$ such that $c_\theta(q^{fb}(\theta),\theta) = c_\theta(q^m(\theta),\theta)$. Hence, $q^m(\theta)$ is a mirror image of $q^{fb}(\theta)$ along $s(\theta)$ with respect to $c_\theta(q,\theta)$.

[80] Obviously, it is enough if the condition holds for all $q \in [q^{fb}(\theta), \bar{q}]$ where $\bar{q}$ is defined as in appendix 4.8.3.

[81] In fact, this also holds true if q^2 in the cost function is replaced by any increasing and convex function.

Proposition 4.4. *Assume that $q^m(\theta)$ is non-decreasing and that there is no distortion at the top.*[82] *Then the optimal solution is below first best and continuous. The optimal solution is strictly increasing at all types where it is below first best if $f(\theta)/(1-F(\theta))$ is non-decreasing and $u_{q\theta} \geq 0$.*

Proof. see appendix

One example for a class of function where $q^m(\theta)$ is increasing are cost functions of the form $c(q,\theta) = \theta q + \phi(q - \alpha\theta) + \gamma(\theta)$ where $\phi(\cdot)$ is a function of which the first three derivatives are positive.[83] Any increasing and concave benefit function $u(q,\theta)$ with $u_{q\theta} = 0$ and $q^{fb}(\theta) > s(\theta)$ yields an increasing $q^m(\theta)$.

Note that in many applications $u_{q\theta} = 0$ will hold. For example, in regulation models, labor market models and monopoly pricing, this property will typically hold because the principal's utility depends only on the decision and the transfer and not directly on the agent's type.

Now it is time to turn to the issue of calculating a solution. In principle, the solution is already described by (4.5), the properties of $\eta(\theta)$ and the necessary conditions C1, C2 and C3. If a non-local incentive constraint binds from a type θ, the three necessary conditions could be used to determine $\hat{\theta}$, $q(\theta)$ and $q(\hat{\theta})$ (assuming that there is a unique solution). If non-local incentive constraints are lax at a type θ, (4.5) can be used to calculate $q(\theta)$ where $\eta(\theta)$ equals $\eta(\hat{\theta}')$ with $\hat{\theta}'$ being defined as the next lower type to which a non-local incentive constraint is binding. While nothing is wrong with this description, it might be burdensome to calculate a solution in this way. Hence, a more structured alternative to obtain a continuous solution might be helpful. This alternative will also give some additional insights into the logic behind the solution. The algorithm is based on the following proposition.

Proposition 4.5. *Define $\Phi^{\eta}(\theta,\hat{\theta})$ as $\Phi(\theta,\hat{\theta})$ under $\tilde{q}(\theta)$ where $\tilde{q}(\theta)$ is derived from*

$$\{u_q(\tilde{q},\theta) - c_q(\tilde{q},\theta)\}f(\theta) + (1 - F(\theta) - \eta)c_{q\theta}(\tilde{q},\theta) = 0.$$

If the incentive constraint binds between θ' and $\hat{\theta}'$ in a continuous solution $q(\theta)$, then

[82] See the following section for a simple sufficient condition for no distortion at the top.

[83] The interpretation of this cost function is that there is a "normal scale" of $\alpha\theta$. Producing above this normal scale is increasingly costly. Type reflects a tradeoff between the size of the normal scale and marginal cost when producing within the normal scale.

$(\theta', \hat{\theta}')$ minimize $\Phi^\eta(\theta, \hat{\theta})$ on $[\hat{\theta}', \theta']$ where $\eta = \eta(\theta') = \eta(\hat{\theta}')$. Furthermore, $\Phi^\eta(\theta', \hat{\theta}') < \Phi^\eta(\theta'', \hat{\theta}'')$ for any $\theta'' > \theta'$ and $\hat{\theta}'' < \hat{\theta}'$.

Proof. see appendix

To get a feeling for this proposition take $\eta = 0$. Then $\tilde{q}(\theta) = q^r(\theta)$. Denote the global minimizer of $\Phi^0(\theta, \hat{\theta})$ by $(\theta^r, \hat{\theta}^r)$. Although a little extra work is needed, the following result follows almost directly from proposition 4.5:

Corollary 4.1. *If the relaxed solution is not implementable, the non-local incentive constraint from θ^r to $\hat{\theta}^r$ will bind in the optimal decision. If one of the two types (both) is interior, his (their) optimal decision is the relaxed decision; i.e. $q(\theta) = q^r(\theta)$ or (and) $q(\hat{\theta}) = q^r(\hat{\theta})$ respectively.*

Proof. see appendix

The proposition then says that a similar logic applies for all pairs $(\theta', \hat{\theta}')$ at which incentive compatibility is binding: One only has to replace $q^r(\theta)$ in the corollary by the corresponding $\tilde{q}(\theta)$. This $\tilde{q}$ is the decision that would result if all types had the same $\eta(\theta)$ and this $\eta(\theta)$ would equal $\eta(\theta')$ in the optimal decision.

The last proposition in connection with theorem 4.1 gives a method for determining $q(\theta)$.

Solve (4.5) for q as a function of type θ and η. Plugging this $q(\theta, \eta)$ into $\Phi(\cdot)$ yields a function $\Phi^\eta(\theta, \hat{\theta})$ which can be minimized over θ and $\hat{\theta}$ yielding $\theta(\eta)$ and $\hat{\theta}(\eta)$ as minimizers. There could be several pairs $(\theta(\eta), \hat{\theta}(\eta))$ locally minimizing $\Phi^\eta(\theta, \hat{\theta})$. Relevant is each pair $(\theta, \hat{\theta})$ (i) that globally minimizes $\Phi^\eta(\cdot)$ on the interval $[\hat{\theta}, \theta]$, (ii) for which no $\Phi^\eta(\cdot)$ minimizer $(\theta', \hat{\theta}')$ with $\theta' > \theta$, $\hat{\theta}' < \hat{\theta}$ and $\Phi^\eta(\theta', \hat{\theta}') < \Phi^\eta(\theta, \hat{\theta})$ exists. For now, assume there is only one such relevant pair.

Under the optimal decision, the constraint will bind from $\theta(\eta)$ to $\hat{\theta}(\eta)$ for all $\eta \in [0, \bar{\eta}]$ where $\bar{\eta}$ is determined by $\Phi^\eta(\theta(\eta), \hat{\theta}(\eta)) = 0$. The optimal decision for types θ where the constraint binds is given by $q(\theta, \eta)$ where η is such that $\theta = \theta(\eta)$. Types for which the constraint does not bind can be sorted into two categories: First, types θ such that non-local incentive constraints do not bind from any type above θ to any type below θ. These types simply have $q(\theta) = q^r(\theta)$. Second, types θ such that the constraint is binding from some $\theta' > \theta$ to some $\hat{\theta}' < \theta$. These types have $\eta(\theta)$ equal to $\eta(\inf\{\theta' : \Phi(\theta', \hat{\theta}') = 0 \text{ with } \theta' > \theta > \hat{\theta}'\})$, i.e. their η is the same as the one of the next lowest

type to which a non-local incentive constraint binds. Their $q(\theta)$ is then $q(\theta, \eta(\theta))$.

One remark on the possibility that several relevant pairs $(\theta(\eta), \hat{\theta}(\eta))$ exist. For example, say there exist the pairs $(\theta_1(\eta), \hat{\theta}_1(\eta))$ and $(\theta_2(\eta), \hat{\theta}_2(\eta))$ both satisfying (i) and (ii) above. The non-local incentive constraint could in this case bind from an interval $[\theta_0, \theta_1]$ to the interval $[\hat{\theta}_1, \hat{\theta}_0]$ as well as from the interval $[\theta_2, \theta_3]$ to the interval $[\hat{\theta}_3, \hat{\theta}_2]$ where $\hat{\theta}_1 < \hat{\theta}_0 < \theta_0 < \theta_1 < \hat{\theta}_3 < \hat{\theta}_2 < \theta_2 < \theta_3$; see figure 4.5 for an illustration. Indeed one has to be a bit more precise in this case: There will be different $\bar{\eta}$ for the two "brackets" of binding incentive constraints. In this case $\eta(\theta)$ will not be single peaked. Hence, the algorithm will then be applied to the two brackets separately and nothing else changes.

A second remark has to be made with regard to bunching. Some types might have an ironed out solution. This solution is then not $q(\theta, \eta(\theta))$ as described above but an ironed out version of it. The condition for determining $\bar{\eta}$, i.e. $\Phi^{\eta}(\theta(\eta), \hat{\theta}(\eta)) = 0$ has to hold for the ironed out decision whenever ironing is relevant. If the monotone hazard rate holds and $u_{q\theta} \geq 0$, one does not have to worry about ironing as long as $\eta \leq 1 - F(\theta(\eta))$: This implies $q(\theta) \leq q^{fb}(\theta)$ for all types for which bunching could have been possible and the decision will be strictly increasing (see the proof of proposition 4.4).

The algorithm is illustrated with a numerical example in the following section.

4.5.4. Distortion at the top

If the non-local incentive constraint binds from $\bar{\theta}$, something unusual can happen. Recall that the necessary condition (C1) might hold with inequality at $\theta = \bar{\theta}$. It is therefore possible that non-local incentive constraints bind from $\bar{\theta}$ to several non-bunched $\hat{\theta}$ even if the solution is continuous. Note that this is impossible for interior types: For a given $q(\theta)$, (C1) and (C2) will uniquely determine $\hat{\theta}$ and $q(\hat{\theta})$.

Now consider the case where the non-local incentive constraint binds not only to several but to a mass of types $\hat{\theta}$ (or to $\underline{\theta}$ as will be shown below). Then the shadow value of the constraint $\eta(\theta)$ will be strictly positive and bounded away from 0 for types slightly below $\bar{\theta}$. Hence, these types have a decision $q(\theta)$ which is at least ε away from their relaxed decision $q^r(\theta)$ for some $\varepsilon > 0$. Obviously, the same has then to apply for $\bar{\theta}$ because of the monotonicity constraint. Put differently, $\eta(\bar{\theta}) > 0$ and therefore $q(\bar{\theta})$ is distorted: There is distortion at the top.

The algorithm described above works also in this situation. The minimizer $\theta(\eta)$ will

then be the boundary type $\bar{\theta}$. The decision of $\bar{\theta}$ and his shadow value are determined by the highest $\hat{\theta}$ to which his non-local incentive constraint binds. At this $\hat{\theta}$ also condition (C1) holds with equality (if $\hat{\theta}$ is above $\underline{\theta}$).

It should be pointed out that distortion at the top is a generic property. Put differently, there will still be distortion at the top if, for example, the distribution of types is slightly perturbed. By proposition 4.5, distortion at the top implies that $\bar{\theta}$ will minimize $\Phi^{\eta}(\theta, \hat{\theta})$ for all $\eta < \tilde{\eta}$ for some $\tilde{\eta} > 0$. $\Phi^{\eta}(\theta, \hat{\theta})$ is continuous in $q(\theta, \eta)$ which in turn is continuous in the density $f(\theta)$. Therefore, $\bar{\theta}$ will remain global minimizer of $\Phi^{\eta}(\theta, \hat{\theta})$ under minor perturbations of the density. Consequently, distortion at the top has to be generic by proposition 4.5.

A natural question is whether there is a sufficient condition for no distortion at the top. Indeed corollary 4.1 allows to formulate such a condition. If $\bar{\theta}$ is not the global minimizer of $\Phi^{r}(\theta, \hat{\theta})$ where $\Phi^{r}(\cdot)$ is $\Phi(\cdot)$ under the relaxed solution $q^{r}(\cdot)$, then non local incentive constraints cannot bind from $\bar{\theta}$. Therefore, the relaxed decision is optimal for $\bar{\theta}$ implying that $q(\bar{\theta}) = q^{fb}(\bar{\theta})$.

Another sufficient condition for no distortion at the top can be formulated using (C1): $\int_0^{q^{fb}(\bar{\theta})} c_{q\theta}(q, \bar{\theta})\, dq \leq 0$ is sufficient since (C1) cannot hold with inequality.

To illustrate the distortion at the top result and also the algorithm introduced in the previous section, consider the following numerical example which is inspired by example 1 in section 4.2.[84]

The cost function is given by $c(q, \theta) = \theta q + \frac{q^2}{\theta} - \frac{\theta}{3}$. The principal's valuation function is $u(q) = \frac{8q}{5}$. Furthermore, I assume that types are distributed on $[1/4, 3/4]$ with density $f(\theta) = 4/5(9\theta - 2)$. Recall from subsection 4.5.3 that with these parameter values the sufficient condition in proposition 4.3 is met. The solution will therefore be continuous.

The first order condition for the relaxed solution is

$$\left(\frac{8}{5} - \theta - \frac{2q}{\theta}\right) * \frac{4}{5}(8\theta - 2) + \frac{33 + 64\theta - 144\theta}{40}\left(1 - \frac{2q}{\theta^2}\right) = 0$$

which leads to the relaxed solution

$$q^{r}(\theta) = \frac{-347\theta^2 + 1660\theta^3 - 2444\theta^4}{330 + 1440\theta^2}.$$

[84] A *Mathematica* notebook with detailed calculations can be found under `https://sites.google.com/site/christophschottmueller/research/webappendices`.

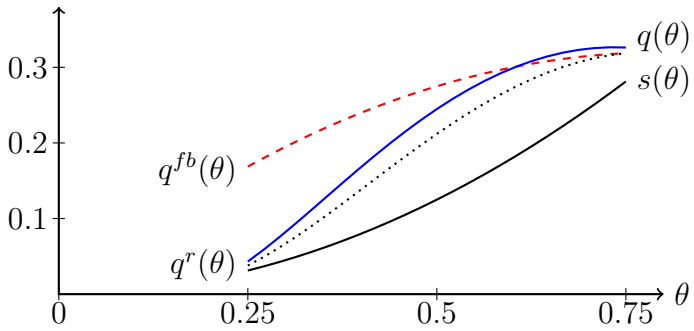


Figure 4.7: numerical example 1

To use the algorithm, $q(\theta, \eta)$ has to be calculated. In this example

$$q(\theta, \eta) = \frac{-2160\theta^4 + 2944\theta^3 - 347\theta^2 - 200\eta\theta^2}{330 - 400\eta + 1440\theta^2}.$$

$\Phi^\eta(\theta, \hat{\theta})$ can be numerically minimized. The result is that $\bar{\theta}$ and $\underline{\theta}$ minimize $\Phi^\eta(\theta, \hat{\theta})$ for all $\eta \leq \bar{\eta} \approx 0.47298$. This means that a non-local incentive constraint is only binding from $\bar{\theta}$ to $\underline{\theta}$ and $\eta(\theta) = 0.47298$ for all types. Consequently, there is distortion at the top and the optimal decision is $q(\theta) = q(\theta, \bar{\eta})$ or

$$q(\theta) = \frac{-\frac{110399}{1250}\theta^2 + \frac{2944}{5}\theta^3 - 432\theta^4}{\frac{17601}{625} + 288\theta^2}.$$

Graphically, figure 4.7 shows that $q(\theta)$ (upper solid line) is above $q^r(\theta)$ (dotted line) for all types and that $q(\theta)$ is above $q^{fb}(\theta)$ (dashed line) for high types.

4.5.5. Stochastic contracts

So far, this paper concentrated on deterministic contracts. Although hardly observed in practice, one could think of stochastic contracts. In the framework of this paper, this would mean that a type θ is assigned a probability distribution over the decision q instead of one deterministic decision $q(\theta)$. The idea behind a stochastic contract is to relax (non-local) incentive constraints. Intuitively, this could work if different types have different degrees of risk aversion. See Rochet (2009) for an example where random contracts are optimal. The following proposition gives a sufficient condition under which deterministic decisions are optimal.

Proposition 4.6. *The optimal decision is deterministic if the assumptions of proposition 4.1, (CVR) and*

$$\frac{\partial \frac{c_{q\theta\theta}}{c_{q\theta}}}{\partial q} \geq 0 \tag{4.6}$$

hold.

Proof. see appendix

Condition (CVR) and (4.6) differ from the conditions for non-stochastic contracts in Maskin and Riley (1984). In Maskin and Riley (1984), only local incentive constraints bind and they bind "downward". It is then shown that assigning the expected decision increases the principal's payoff and relaxes local incentive constraints if risk aversion is decreasing in type. Decreasing risk aversion is therefore a sufficient condition for the optimality of deterministic contracts. This reasoning is flawed in case non-local incentive constraints are binding: Assigning the expected decision decreases the slope of the rent function $\pi(\theta)$ because $-c_\theta$ is convex in q. Hence, profit differences between θ and $\hat{\theta}$ are smaller under the expected decision compared to the stochastic contract, i.e. non-local incentive constraints are harder to satisfy.

Proposition 4.6 takes therefore another way which is also taken in Jullien (2000). When rewriting the principal's optimization problem as an optimization over rent profiles π (instead of over decision q) condition (CVR) ensures that the resulting program is concave. Condition (4.6) ensures that the set of implementable utility profiles is convex. These two properties imply that a stochastic decision is worse for the principal than a deterministic decision implementing the same utility profile. The conditions of proposition 4.1 allow to focus on decisions above $s(\theta)$ which correspond to monotone solutions.

4.6. Discussion

I want to discuss the assumptions on third derivatives, i.e. $c_{qq\theta} < 0$ and $c_{q\theta\theta} > 0$. The fact that these derivatives do not change sign ensures that the cross derivative $c_{q\theta}$ changes sign only once for any given θ (or q). While this property is admittedly important for the analysis, it is immaterial which sign the third derivatives have (as long as the sign is the same for all relevant decisions and types). To illustrate this (and also to show an example where the monotonicity constraint binds) consider the following version of example 2:[85] Types are distributed uniformly on $[2, 3]$ and the principal's objective is

[85] A *Mathematica* notebook with detailed calculations can be found under `https://sites.google.com/site/christophschottmueller/research/webappendices`.

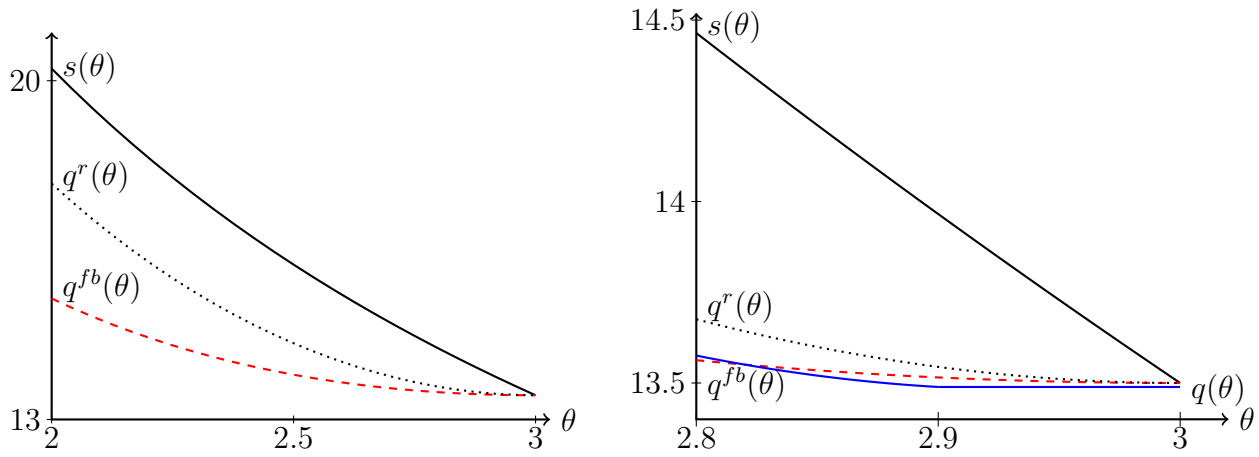


Figure 4.8: numerical example 2

the expected value of $q(\theta) - t(\theta)$. The agent's utility is given by

$$\pi(q, \theta) = t(\theta) - \frac{(q - \theta/\sigma)^2}{\theta^2} + \gamma(3 - \theta).$$

Here the parameter values $\sigma = 27$ and $\gamma = 12$ are used. In this case, third derivatives have the following signs in the relevant range of the decision: $c_{qq\theta} < 0$ and $c_{q\theta\theta} < 0$. Consequently, the sign switching decision $s(\theta)$ is downward sloping. As depicted in figure 4.8a, first best decision and relaxed decision are also downward sloping.

Although the example looks different on first sight, it is equivalent to the model of the main text and all results apply accordingly. It turns out that also in this example $(\bar{\theta}, \underline{\theta})$ minimize $\Phi^\eta(\theta, \hat{\theta})$ and therefore only the non-local incentive constraint from the highest to the lowest type is binding. However, the monotonicity constraint is binding for the highest types. For each $q(\theta, \eta)$, the optimal bunching interval $[\theta_s(\eta), \bar{\theta}]$ is determined by the condition

$$\int_{\theta_s(\eta)}^{\bar{\theta}} [u_q(q(\theta, \eta), \theta) - c_q(q(\theta, \eta), \theta)] f(\theta) + (1 - F(\theta) - \eta(\theta)) c_{q\theta}(q(\theta, \eta), \theta)\, d\theta = 0.$$

Here, $\bar{\eta}$ turns out to be approximately 0.18 and the solution for the highest types is depicted in figure 4.8b. The solution exhibits bunching of types in $[2.9, 3]$.

It should be pointed out that the results in this paper do not necessarily hold if one of the third derivatives $c_{qq\theta}$ or $c_{q\theta\theta}$ changes sign in the relevant range of q. To see this, imagine that in figure 4.4a there was a second $s(\theta)$ function where the $c_{q\theta}$ changes sign again and assume that this second $s(\theta)$ was above $q(\theta)$ but intersecting the shaded area in figure 4.4a. In this case, lemma 4.1 would not necessarily hold and the results building

on lemma 4.1 might also be affected. For example, theorem 4.1 would have to be refined: The properties for η_θ at types from and to which non-local incentive constraints bind will only hold at types $\theta > \hat{\theta}$ where non-local incentive constraints are only downward binding.

4.7. Conclusion

This paper characterizes monotone solutions in a screening environment where single crossing is violated. Although the model restricts itself to a one time violation of single crossing, the main effects of a violation of single crossing can be illustrated. Non-local incentive constraints can become binding. The distortion caused by non-locally binding incentive constraints can counteract the normal rent extraction distortion. Therefore, the solution can be partly above as well as below the first best decision. There can be distortion at the top if non local incentive constraints are binding from the top type to a mass of types (or the lowest type). Furthermore, sufficient conditions for monotonicity and continuity are provided and an algorithm for determining such a continuous, monotone solution is proposed.

Possible applications can be found in various fields of economics. While the paper uses the notation of a regulation or procurement setting, the same model is applicable, for example, in models of labor, insurance, monopoly pricing or optimal taxation. The characterization of continuous and monotone solutions is relatively simple and reasonable classes of functions satisfy sufficient conditions for falling into this class of solutions.

I conclude with some immediate implications of the qualitative results in this paper. In optimal taxation models where single crossing is violated, the distortion above first best result would correspond to negative marginal tax rates. Earned income tax credit schemes often lead to negative marginal tax rates for low income workers. Previous theoretical arguments based on externalities or general equilibrium effects as in Stiglitz (1982) could only explain negative marginal tax rates for productive types but are less applicable for low income workers. Non-local incentive constraints binding to low ability types could however lead to the observed pattern.[86]

[86] An alternative explanation presented by Chone and Laroque (2010) is based on the possibility that participation constraints might be binding not for the lowest type but for types of higher ability.

Negative marginal tax rates for top incomes can be rationalized because of the distortion at the top result. Note that distortion at the top is always in an "unusual" direction, i.e. above first best. The rough intuition is that subsidizing productive types to work more increases their rent and therefore relaxes their incentive compatibility constraint.

Overinsurance can be optimal in insurance models where single crossing is violated. This gives an alternative explanation for so called "Cadillac" health insurance plans. While the political debate focuses on viewing them as (insufficiently taxed) part of a compensation package, screening by insurers with market power could also explain parts of the phenomenon.

In Martimort and Stole (2009) the ordering of first best quantities and the competitive menu under substitutes is no longer clear cut if one considers the cases without single crossing. Put differently, firms using non-linear pricing might optimally offer packages which lead to overconsumption of the good. Telecommunication might be an example for this: Consumers often buy packages where an additional unit of calling (or internet use) is for free. If the marginal costs of the provider are only ε above zero, such a price scheme will lead to consumption above the socially optimal consumption. Overconsumption in terms of quality even at the top is already described in Dupuit (1849) following his explanation of downward distortion of quality in the third and second class he continues:

> And it is again for the same reason that the companies, having proved almost cruel to third-class passengers and mean to second-class ones, have become lavish in dealing with first class passengers. Having refused the poor what is necessary, they give the rich what is superfluous.

The last point also relates to the empirical literature on non-linear pricing. The standard no distortion at the top result allows for a simple way to identify (constant) marginal costs: For the highest type, marginal tariff and marginal costs have to be equal. Miravete and Röller (2004), for example, use this condition to recover marginal costs. As there can be distortion at the top without single crossing, this possibility is no longer available. This is especially relevant for competitive non-linear pricing since Martimort and Stole (2009) show that single crossing can be violated even with standard preferences if there are competing firms.

4.8. Appendix

4.8.1. Variational condition

In Araujo and Moreira (2010), it always holds that $q(\theta) = q(\hat{\theta})$ whenever $\Phi(\theta, \hat{\theta}) = 0$.[87] Consequently, (C1) does not play a role. Starting from (C2), they derive the following condition (with $q = q(\theta) = q(\hat{\theta})$):

$$\frac{u_q(q,\theta) - c_q(q,\theta) + \frac{1-F(\theta)}{f(\theta)} c_{q\theta}(q,\theta)}{c_{q\theta}(q,\theta)} f(\theta) = \frac{u_q(q,\hat{\theta}) - c_q(q,\hat{\theta}) + \frac{1-F(\hat{\theta})}{f(\hat{\theta})} c_{q\theta}(q,\hat{\theta})}{c_{q\theta}(q,\hat{\theta})} f(\hat{\theta}) \quad (4.7)$$

To derive a similar condition for $q(\theta) \neq q(\hat{\theta})$ take θ and $\hat{\theta}$ such that $c_q(q(\hat{\theta}), \hat{\theta}) = c_q(q(\hat{\theta}), \theta)$, $c_\theta(q(\theta), \theta) = c_\theta(q(\hat{\theta}), \theta)$, $\Phi(\theta, \hat{\theta}) = 0$ and assume that $q(\cdot)$ is strictly monotone and continuous at θ and $\hat{\theta}$.

Given θ and $q(\theta)$, the equation $c_\theta(q(\theta), \theta) = c_\theta(q(\hat{\theta}), \theta)$ pins down a decision $q(\hat{\theta})$ where incentive compatibility could be binding. Given this $q(\hat{\theta})$ as well as θ and $q(\theta)$, the equation $c_q(q(\hat{\theta}), \hat{\theta}) = c_q(q(\hat{\theta}), \theta)$ determines $\hat{\theta}$. Therefore, the critical $\hat{\theta}$ can be written as a function of θ and $q(\theta)$, i.e. $\hat{\theta} = \phi(\theta, q(\theta))$.

Differentiating the two conditions, the partial derivatives ϕ_θ and ϕ_q can be obtained as

$$\begin{aligned}
\phi_\theta(\theta, q) &= \frac{c_{q\theta}(\hat{q}, \theta)}{c_{q\theta}(\hat{q}, \hat{\theta})} + \frac{(c_{qq}(\hat{q}, \theta) - c_{qq}(\hat{q}, \hat{\theta}))(c_{\theta\theta}(q, \theta) - c_{\theta\theta}(\hat{q}, \theta))}{c_{q\theta}(\hat{q}, \hat{\theta}) c_{q\theta}(\hat{q}, \theta)} \\
\phi_q(\theta, q) &= \frac{c_{q\theta}(q, \theta)[c_{qq}(\hat{q}, \theta) - c_{qq}(\hat{q}, \hat{\theta})]}{c_{q\theta}(\hat{q}, \hat{\theta}) c_{q\theta}(\hat{q}, \theta)}
\end{aligned}$$

where $\hat{q} = q(\hat{\theta})$ and $q = q(\theta)$.

Denote by h an admissible perturbation of the optimal solution q^* on some interval $[\theta_1, \theta_2]$, i.e. $h(\theta_1) = h(\theta_2) = 0$. Admissibility implies that if the incentive constraint binds from θ to $\hat{\theta}$, then $\hat{\theta} = \phi(\theta, q(\theta))$.[88]

The idea of the variational argument is the following: I want to derive a necessary condition for a type θ such that $\Phi(\theta, \hat{\theta}) = 0$ for some $\hat{\theta}$. To do so, it is assumed that also under the perturbed decision the incentive constraint is binding for θ and some (other) $\hat{\theta}$. The type $\hat{\theta}$ to which the non-local incentive constraint binds depends on

[87] The variational condition of this appendix was also derived in the working paper version Araujo and Moreira (2001).

[88] Furthermore, admissibility requires monotonicity.

the perturbation and is given by $\phi(\theta, q(\theta))$. The way one should think about it is that incentive compatibility is binding from each $\theta \in [\theta_1, \theta_2]$ to some $\hat{\theta}$ in some interval $[\hat{\theta}_1, \hat{\theta}_2]$.[89] The specific type $\hat{\theta}$ to which a non-local incentive constraint binds from a given θ depends on the perturbation h.

For brevity, I denote in the remainder of this section the optimal solution by $q^*(\theta)$ and the perturbed solution by $q(\theta) = q^*(\theta) + \varepsilon h(\theta)$. Hence the part of the principal's objective function affected by the perturbation can be written as[90]

$$\begin{aligned} G(\varepsilon) &= \int_{\theta_1}^{\theta_2} g(q(\theta), \theta)\, d\theta + \int_{\phi(\theta_2, q(\theta_2))}^{\phi(\theta_1, q(\theta_1))} g(q(\theta), \theta)\, d\theta \\ &= \int_{\theta_1}^{\theta_2} \{g(q(\theta), \theta) - g(\hat{q}(\theta, q(\theta)), \phi(\theta, q(\theta)))\, [\phi_q(q(\theta), \theta) q_\theta(\theta) + \phi_\theta(q(\theta), \theta)]\}\, d\theta \end{aligned} \tag{4.8}$$

where $g(q(\theta), \theta) = \left[u(q(\theta), \theta) - c(q(\theta), \theta) + \frac{1-F(\theta)}{f(\theta)} c_\theta(q(\theta), \theta)\right] f(\theta)$ is the virtual valuation weighted by the density. The second line is a normal change of variables where $\hat{q}(\theta, q)$ denotes the $\hat{q}$ solving $c_\theta(q, \theta) = c_\theta(\hat{q}, \theta)$ with $q \neq \hat{q}$. Note that $\partial \hat{q}/\partial q = c_{q\theta}(q, \theta)/c_{q\theta}(\hat{q}, \theta)$.

Differentiating (4.8) gives

$$G'(0) = \int_{\theta_1}^{\theta_2} \{g_q h - \hat{g}((\phi_{qq} q_\theta^* + \phi_{q\theta})h + \phi_q h_\theta) - (\hat{g}_q \hat{q}_q + \hat{g}_\theta \phi_q)(\phi_q q_\theta^* + \phi_\theta)h\}\, d\theta = 0$$

where arguments are omitted and a hat denotes evaluation at $(\hat{\theta}, q^*(\hat{\theta}))$. Integrating $\int_{\theta_1}^{\theta_2} (\hat{g}\phi_q) h_\theta\, d\theta$ by parts and substituting yields for the previous equation

$$\int_{\theta_1}^{\theta_2} \{g_q - \hat{g}_q \hat{q}_q \phi_\theta + \hat{g}_q \hat{q}_\theta \phi_q\}\, h\, d\theta = \int_{\theta_1}^{\theta_2} \left\{ g_q - \hat{g}_q \frac{c_{q\theta}(q(\theta), \theta)}{c_{q\theta}(q(\hat{\theta}), \hat{\theta})} \right\} h\, d\theta = 0.$$

As h was arbitrary, the following condition has to hold at optimum:

$$g_q(q(\theta), \theta) = g_q(q(\hat{\theta}), \hat{\theta}) \frac{c_{q\theta}(q(\theta), \theta)}{c_{q\theta}(q(\hat{\theta}), \hat{\theta})} \tag{C3'}$$

This is condition (C3). For $q(\theta) = q(\hat{\theta})$, (C3') boils down to (4.7).

4.8.2. Proofs

Proof of proposition 4.1: First, it is shown that the principal's payoff is higher under $q^c(\theta)$ than under $q(\theta)$: The principal maximizes expectation of $u(q, \theta) - c(q, \theta) + (1 - F(\theta))/f(\theta) c_\theta(q, \theta)$. If $q^s(q, \theta) \leq q^r(\theta)$, the principal's objective increases due to the

[89] As it turns out, this is indeed the typical structure of a continuous solution, see lemma 4.4.

[90] It follows from lemma 4.2 that $\phi(\theta_1, q(\theta_1)) > \phi(\theta_2, q(\theta_2))$.

change because of the concavity of (RP) and $q^r(\theta) > s(\theta)$. If $q^s(q(\theta),\theta) > q^{fb}(\theta)$, then the same conclusion follows from $q^v(q(\theta),\theta) \geq q^s(q(\theta),\theta) > q^r(\theta)$ and the concavity of (RP).

Second, the changed decision $q^c(\theta)$ is monotonically increasing: From local incentive compatibility $q(\theta)$ was already increasing wherever it was above $s(\theta)$. At types with $q(\theta) < s(\theta)$ the decision $q(\theta)$ had to be decreasing because of local incentive compatibility. But then $q^s(q(\theta),\theta)$ is clearly increasing in θ for these types because of $c_{q\theta\theta} > 0$. This leaves types at which $q(\theta)$ jumped discontinuously over $s(\theta)$. But at these jump types local incentive compatibility required $c_\theta(q^-(\theta),\theta) - c_\theta(q^+(\theta),\theta) \geq 0$ at downwards jumps (and the converse inequality at upwards jumps) across $s(\theta)$. This implies that also at jump points of $q(\theta)$ monotonicity of $q^c(\theta)$ is guaranteed.

Third, the changed decision $q^c(\theta)$ is incentive compatible: Since $q^c(\theta)$ is monotonically increasing, only downward misrepresentation has to be considered (see lemma 4.1). Note that the profit function $\pi(\theta)$ was not affected by the change from $q(\theta)$ to $q^c(\theta)$ because of the definition of $q^s(\theta)$ and $\pi_\theta(\theta) = -c_\theta(q(\theta),\theta)$ by local incentive compatibility. Therefore, one has only to check whether any type wants to misrepresent as a lower type $\hat{\theta}$ at which $q(\hat{\theta}) < s(\hat{\theta})$. Since $\pi(\theta)$ is unchanged, one can write incentive compatibility under the changed decision as

$$\begin{aligned}\Phi^c(\theta,\hat{\theta}) &= -\int_{\hat{\theta}}^{\theta}\int_{q^c(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt = -\int_{\hat{\theta}}^{\theta}\int_{q^c(\hat{\theta})}^{q(\hat{\theta})} c_{q\theta}(q,t)\,dq\,dt - \int_{\hat{\theta}}^{\theta}\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt \\ &= \int_{\hat{\theta}}^{\theta}\int_{q(\hat{\theta})}^{q^c(\hat{\theta})} c_{q\theta}(q,t)\,dq\,dt + \Phi(\theta,\hat{\theta}) > 0\end{aligned}$$

where the inequality follows from $\int_{q(\hat{\theta})}^{q^c(\hat{\theta})} c_{q\theta}(q,\hat{\theta})\,dq = 0$ by the definition of $q^s(\cdot)$ and $c_{q\theta\theta} > 0$. *Q.E.D.*

Proof of lemma 4.3: First, it is shown that there cannot be a discontinuity at $\hat{\theta}$. Take a type $\hat{\theta}$ to which non-local incentive constraint is binding from some type θ. Suppose that $q(\cdot)$ is discontinuous at $\hat{\theta}$, i.e. $q^-(\hat{\theta}) < q^+(\hat{\theta})$ by local incentive compatibility (monotonicity). Binding incentive constraint means that either (i) $\int_{\hat{\theta}}^{\theta}\int_{q^-(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt = 0$ or (ii) $\int_{\hat{\theta}}^{\theta}\int_{q^+(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt = 0$ or (iii) $q^-(\hat{\theta}) < q(\hat{\theta}) < q^+(\hat{\theta})$ and $\int_{\hat{\theta}}^{\theta}\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt = 0$.[91]

[91] Case (iii) does not imply that all/several decisions between $q^-(\hat{\theta})$ and $q^+(\hat{\theta})$ are offered. Only one decision is offered for each type and at a discontinuity this decision might be strictly between the two

In case (i) it must hold that $\int_{\hat{\theta}}^{\theta} c_{q\theta}(q^-(\hat{\theta}),t)\,dt \leq 0$ which is just (C2) adapted to apply for a right hand side discontinuity, i.e. if this did not hold incentive compatibility would be violated for θ and $\hat{\theta}-\varepsilon$. But then $\int_{\hat{\theta}}^{\theta}\int_{q^-(\hat{\theta})}^{q^+(\hat{\theta})} c_{q\theta}(q,t)\,dq\,dt < 0$ from $c_{qq\theta} < 0$. Hence, $\Phi(\theta,\hat{\theta}^+) = \Phi(\theta,\hat{\theta}^-) + \int_{\hat{\theta}}^{\theta}\int_{q^-(\hat{\theta})}^{q^+(\hat{\theta})} c_{q\theta}(q,t)\,dq\,dt < 0$ as $\Phi(\theta,\hat{\theta}^-) = 0$ by assumption. Hence, incentive compatibility is violated from θ to types slightly above $\hat{\theta}$. This is the desired contradiction.

In case (ii) it must hold that $\int_{\hat{\theta}}^{\theta} c_{q\theta}(q^+(\hat{\theta}),t)\,dt \geq 0$. But then $\int_{\hat{\theta}}^{\theta}\int_{q^-(\hat{\theta})}^{q^+(\hat{\theta})} c_{q\theta}(q,t)\,dq\,dt > 0$ from $c_{qq\theta} < 0$. Consequently, $\Phi(\theta,\hat{\theta}^-) = \Phi(\theta,\hat{\theta}^+) - \int_{\hat{\theta}}^{\theta}\int_{q^-(\hat{\theta})}^{q^+(\hat{\theta})} c_{q\theta}(q,t)\,dq\,dt < 0$ and therefore incentive compatibility is violated from θ to types slightly below $\hat{\theta}$.

In case (iii) the same arguments as in case (i) apply if $\int_{\hat{\theta}}^{\theta} c_{q\theta}(q(\hat{\theta}),t)\,dt \leq 0$ while the same arguments as in case (ii) apply if $\int_{\hat{\theta}}^{\theta} c_{q\theta}(q(\hat{\theta}),t)\,dt > 0$.

Second, it is shown that $\theta < \bar{\theta}$ cannot be bunched with some type θ' if $q(\cdot)$ is continuous at θ. Suppose θ and θ' were bunched on q^b (and by monotonicity all types in between them are as well) and suppose for now $\theta < \theta'$. But then $\Phi(\theta',\hat{\theta}) < 0$ and ic is violated as

$$\begin{aligned}\Phi(\theta',\hat{\theta}) &= -\int_{\hat{\theta}}^{\theta'}\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(s,t)\,ds\,dt = -\int_{\hat{\theta}}^{\theta}\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(s,t)\,ds\,dt - \int_{\theta}^{\theta'}\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(s,t)\,ds\,dt \\ &= \Phi(\theta,\hat{\theta}) - \int_{\theta}^{\theta'}\int_{q(\hat{\theta})}^{q^b} c_{q\theta}(s,t)\,ds\,dt < 0\end{aligned}$$

where the last inequality follows from (C1) and $c_{q\theta\theta} > 0$.

Now suppose $\theta > \theta'$ and both types are bunched. From condition (C1) for $\theta < \bar{\theta}$ and $c_{q\theta\theta} > 0$ it follows that $\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq < 0$ for every $t \in (\theta-\varepsilon,\theta)$. But then $\Phi(\theta-\varepsilon,\hat{\theta}) = \Phi(\theta,\hat{\theta}) + \int_{\theta-\varepsilon}^{\theta}\int_{q(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt < 0$, so incentive compatibility would be violated. *Q.E.D.*

Proof of proposition 4.2: Suppose $q(\theta) < q^r(\theta)$ for some types. Since local incentive compatibility does not allow downward jumps, $q(\theta)$ has to be strictly below $q^r(\theta)$ for a mass of types. Consider changing this 'optimal' decision to $q^*(\theta)$ where $q^*(\theta) = max\{q(\theta), q^r(\theta)\}$. Transfers $t^*(\theta)$ are determined such that $\pi(\underline{\theta}) = 0$ and $\pi_\theta(\theta) = -c_\theta(q^*(\theta),\theta)$.

By the definition of $q^r(\theta)$, this change will increase the principal's expected payoff.

limits (case (iii)).

It remains to check incentive compatibility, i.e

$$\Phi^*(\theta,\hat{\theta}) = -\int_{\hat{\theta}}^{\theta}\int_{q^*(\hat{\theta})}^{q^*(t)} c_{q\theta}(q,t)\,dq\,dt \geq 0$$

for arbitrary types θ and $\hat{\theta} < \theta$. If $q^*(\hat{\theta}) = q(\hat{\theta})$, incentive compatibility follows from $q^*(t) \geq q(t)$ and as $q(t) \geq s(t)$ the corresponding 'additional' $c(q,t)$ are negative.

If $q^*(\hat{\theta}) > q(\hat{\theta})$ (and therefore $q^*(\hat{\theta}) = q^r(\hat{\theta})$), there are three possibilities: (i) There exists a type $\theta' \in (\hat{\theta},\theta)$ with $q(\theta') = q^*(\hat{\theta})$, (ii) all types $\theta' \in (\hat{\theta},\theta)$ have $q(\theta') < q^*(\hat{\theta})$ and (iii) there are types $\theta' \in (\hat{\theta},\theta)$ with $q(\theta') > q^*(\hat{\theta})$ but no type θ' with $q(\theta') = q^*(\hat{\theta})$, hence $q(\cdot)$ is discontinuous[92].

If (i), then $\Phi(\theta,\theta') \geq 0$ implies incentive compatibility as $\Phi^*(\theta,\hat{\theta}) > \Phi(\theta,\theta')$. In case (ii) $q^*(\hat{\theta})$ has to be above $q(\theta')$ for all $\theta' \in (\hat{\theta},\theta)$. But since $q(\theta') > s(\theta')$ for all these types it follows that $q^*(\hat{\theta}) > s(\theta)$ and therefore incentive compatibility is trivially satisfied.

In case (iii) define $\theta' = sup\{t \in (\hat{\theta},\theta) :\ q(t) < q^*(\hat{\theta})\}$ that is θ' is the jump point. Incentive compatibility between θ and θ' implies $\int_{\theta'}^{\theta}\int_{q^-(\theta')}^{q(t)} c_{q\theta}(q,t)\,dq\,dt \leq 0$ as well as $\int_{\theta'}^{\theta}\int_{q^+(\theta')}^{q(t)} c_{q\theta}(q,t)\,dq\,dt \leq 0$ where $q^-(\theta')$ denotes the limit of $q(t)$ as $t \to \theta'$ from below. From $c_{qq\theta} < 0$ and $q^-(\theta') < q^*(\hat{\theta}) < q^+(\theta')$, it follows that $\int_{\theta'}^{\theta}\int_{q^*(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt \leq 0$. But as $\Phi^*(\theta,\hat{\theta}) > -\int_{\theta'}^{\theta}\int_{q^*(\hat{\theta})}^{q(t)} c_{q\theta}(q,t)\,dq\,dt \geq 0$ incentive compatibility is satisfied. *Q.E.D.*

Proof of theorem 4.1: Note that even if the theorem was not true one could still define a function $\eta(\theta)$ by rearranging (4.5). What one has to show are the properties of this function. $\eta(\theta) \geq 0$ follows immediately from proposition 4.2 and the fact that the left hand side of (4.5) is decreasing in q.

Next turn the property that $\eta(\theta)$ is constant on an interval of types on which non local incentive constraints are lax. Suppose to the contrary that $\eta(\theta)$ is not constant. In particular, suppose $\eta(\theta)$ was increasing on some interval $[\theta_1,\theta_3]$ where non-local ic is lax for all $\theta \in [\theta_1,\theta_3]$. Denote by θ_2 some interior type of the interval. For each $\theta \in [\theta_2,\theta_2+\varepsilon]$ define a corresponding type $\theta' \in [\theta_2-\varepsilon,\theta_2]$ by $\theta' = \theta_2 - (\theta-\theta_2)$ for some small $\varepsilon > 0$. I will show that one can change such a decision on $[\theta_2-\varepsilon,\theta_2+\varepsilon]$ in a way which increases the principal's payoff (while keeping incentive compatibility). This contradicts the optimality of $q(\theta)$.

Consider a changed decision $q^c(\cdot)$ such that (i) $q^c(\theta) > q(\theta)$ on $[\theta_2-\varepsilon,\theta_2)$, (ii) $q^c(\theta) \leq$

[92]Given that solutions in Araujo and Moreira (2010) display sometimes discontinuities, one cannot totally exclude this possibility.

$q(\theta)$ on $[\theta_2, \theta_2+\varepsilon]$, (iii) for corresponding types θ and θ' it holds that $\int_{q(\theta')}^{q^c(\theta')} c_{q\theta}(q,\theta')\,dq = -\int_{q(\theta)}^{q^c(\theta)} c_{q\theta}(q,\theta)\,dq$ and (iv) $q^c_\theta(\theta) \geq 0$ on $[\theta_2-\varepsilon, \theta_2+\varepsilon]$. The changed decision will therefore display upwards jumps at $\theta_2-\varepsilon$ and $\theta_2+\varepsilon$. For small changes in q, (iii) can be written as $\delta(\theta')c_{q\theta}(q(\theta'),\theta') = -\delta(\theta)c_{q\theta}(q(\theta),\theta)$ where $\delta(\theta) = q^c(\theta) - q(\theta)$. This in turn can be written as $\delta(\theta') = -\delta(\theta)k(\theta)$ where $k(\theta)$ is defined as $\frac{c_{q\theta}(q(\theta),\theta)}{c_{q\theta}(q(\theta'(\theta)),\theta'(\theta))}$.

Before proceeding, let me show that a function $q^c(\theta)$ satisfying (i)-(iv) exists. Note that $k(\theta_2) = 1$ and that–due to the differentiability and continuity assumptions on $c(\cdot)$ and the monotonicity of $q(\theta)$–the function $k(\theta)$ is continuously differentiable almost everywhere.[93] First, consider the case where $k^+_\theta(\theta_2) < 0$. Then it is feasible to set $q^c(\theta) = q(\theta_2)$ for types $\theta \in [\theta_2, \theta_2+\varepsilon]$ if $\varepsilon > 0$ is chosen small enough. Feasibility means that determining $q(\theta')$ by $\delta(\theta') = -\delta(\theta)k(\theta)$ will satisfy all conditions especially (iv). Feasibility of $q^c(\theta) = q(\theta_2)$ for $\theta \in [\theta_2, \theta_2+\varepsilon]$ and monotonicity of $q(\theta)$ imply that $q^{c*} = \alpha q^c(\theta) + (1-\alpha)q(\theta)$ is also feasible. The effect of a marginal change of q is the effect changing $q(\cdot)$ to $q^{c*}(\cdot)$ as $\alpha \to 0$.

Second, consider $k_\theta(\theta_2)^+ > 0$. By the same argument, it is feasible to bunch types $\theta \in [\theta_2-\varepsilon, \theta_2]$ on $q(\theta_2)$ and the remaining argument goes through analogously. Obviously, the third case $k^+_\theta(\theta_2) = 0$ is analogous to either the first or the second case (depending on the second derivative).

The effect of a marginal change on the principal's objective is

$$\int_{\theta_2-\varepsilon}^{\theta_2+\varepsilon} \{(u_q(q(\theta),\theta) - c_q(q(\theta),\theta))f(\theta) + (1-F(\theta))c_{q\theta}(q(\theta),\theta)\}\,\delta(\theta)\,d\theta$$
$$= \int_{\theta_2-\varepsilon}^{\theta_2+\varepsilon} \eta(\theta)c_{q\theta}(q(\theta),\theta)\delta(\theta)\,d\theta = \int_{\theta_2}^{\theta_2+\varepsilon} \delta(\theta)c_{q\theta}(q(\theta),\theta)[\eta(\theta) - \eta(\theta'(\theta))] > 0$$

where the last inequality follows from $\delta(\theta) \leq 0$ for $\theta \in [\theta_2, \theta_2+\varepsilon]$ and $\eta_\theta(\theta) > 0$. Hence, the principal's objective increases. Due to (iii) incentive compatibility is still satisfied. This contradicts the optimality of $q(\theta)$.

A similar argument can be made when $\eta(\theta)$ is decreasing almost everywhere on some interval $[\theta_1, \theta_3]$ where non-local ic is lax. The only difference is that (i) and (ii) are substituted by (i) $q^c(\theta) < q(\theta)$ on $[\theta_2-\varepsilon, \theta_2)$, (ii) $q^c(\theta) \geq q(\theta)$ on $[\theta_2, \theta_2+\varepsilon]$. The argument for existence is then that for $k_\theta(\theta_2) < 0$ one can choose a $\theta_2+\varepsilon$ such that

[93] Note that a feasible $q^c(\theta)$ exists even around types θ_2 where $q(\theta)$ is discontinuous: Whether bunching types $[\theta_2-\varepsilon, \theta_2)$ on $q^-(\theta_2)$ or bunching types $(\theta_2, \theta_2+\varepsilon]$ on $q^+(\theta_2)$ is feasible is then decided by $k^+_\theta(\theta_2)$ just as in the text.

setting $q^c(\theta) = q(\theta_2 + \varepsilon)$ for all $\theta \in [\theta_2, \theta_2 + \varepsilon]$ is feasible. Everything else goes through accordingly.

Hence, $\eta(\theta)$ is constant on all intervals on which non-local incentive constraints do not bind.[94]

To see that $\eta(\theta)$ is non-decreasing at types $\hat{\theta}$ to which a non-local incentive constraint is binding one can use the same steps as above for types where non-local incentive constraints were lax. The key insight is that such a change is feasible due to the structure given by lemma 4.1 and lemma 4.2 (see also figure 4.5): Increasing q for slightly higher types than $\hat{\theta}$ (and reducing for slightly lower types than $\hat{\theta}$) will relax (or not affect) binding non-local incentive constraints because these constraints are downward binding and not overlapping.

The argument why $\eta(\theta)$ is non-increasing at types θ from which non-local incentive constraints bind is also equivalent to the one above. The key with respect to feasibility is now that reducing q for types slightly below θ (and increasing for types slightly above θ) will again relax (or not affect) binding non-local incentive constraints because these constraints are downward binding.

Now turn to $\eta(\bar{\theta}) = 0$ (and therefore $q(\bar{\theta}) = q^{fb}(\bar{\theta})$) whenever no non-local incentive constraint is binding from $\bar{\theta}$. Clearly, $q(\bar{\theta})$ does not affect non-local incentive constraints of other types, see figure 4.2b for an illustration. Consequently, the principal's payoff is maximized by setting $q(\bar{\theta}) = q^r(\bar{\theta})$. The only thing to show is that the monotonicity constraint is not binding at $\bar{\theta}$. Suppose to the contrary that types $[\theta', \bar{\theta}]$ were bunched on $q^b > q^{fb}(\bar{\theta})$. By lemma 4.3, non-local incentive constraints cannot be binding for types in $(\theta', \bar{\theta}]$. First, note that $q(\theta)$ has to be continuous at θ' as otherwise the principal's payoff could be increased by reducing q^b. Therefore–by the same argument as in the proof of lemma 4.3–non-local incentive constraints cannot bind from types $[\theta' - \varepsilon, \theta']$ for some small $\varepsilon > 0$. Given that $q(\theta) > q^{fb}(\bar{\theta}) > q^r(\theta)$ for all $\theta \in [\theta' - \varepsilon, \bar{\theta})$, the principal's payoff could be increased by changing $q(\theta)$ to $q(\theta' - \varepsilon)$ for all $\theta \in [\theta' - \varepsilon, \bar{\theta}]$. This contradicts the optimality of $q(\theta)$.

The part that $\eta(\underline{\theta}) = 0$ if no non-local incentive constraint is binding to $\underline{\theta}$ is even

[94] Note that $\eta(\theta)$ cannot be different for isolated types in such an interval: This would, by (4.5) and the continuity of the derivatives of $c(\cdot)$, lead to $q(\theta)$ being discontinuous at isolated points. Such a discontinuity, however, violates local incentive compatibility.

simpler: Reducing $q(\theta)$ to $q^r(\theta)$ cannot violate the monotonicity constraint as $q(\theta) \geq q^r(\theta) \geq q^r(\underline{\theta})$ by proposition 4.2. *Q.E.D.*

Proof of lemma 4.4: I proof the stronger statement, i.e. non-local incentive constraints do not only bind at isolated interior types. The proof is by contradiction.

Suppose, non-local incentive constraints bound only from isolated interior types. Denote by θ' the supremum of all types with $\eta(\theta) > 0$, i.e. $\theta' = \sup\{\theta : \ \eta(\theta) > 0\}$. By theorem 4.1, a non-local incentive constraint is binding from θ' and $\eta(\theta) = 0$ for all $\theta > \theta'$.[95] As the set of types from which non-local incentive constraints bind consists only of isolated types, there exists an $\varepsilon > 0$ such that non-local incentive constraints are lax for all $\theta \in (\theta' - \varepsilon, \theta')$. By theorem 4.1, $\eta(\theta)$ is constant on $(\theta' - \varepsilon, \theta')$ and by the definition of θ' there has to be a discontinuity in $\eta(\theta)$ at θ', i.e. $\eta^-(\theta') > \eta^+(\theta') = 0$. The definition of $\eta(\theta)$ in (4.5) implies then that $q^-(\theta') > q^+(\theta')$. But this violates the monotonicity constraint. Hence, θ' cannot be isolated in the set of types from which non-local incentive constraints bind.

Similarly, take $\hat{\theta}' = \inf(\hat{\theta} : \ \eta(\hat{\theta}) > 0)$. It holds that $\eta(\theta) = 0$ for all $\theta < \hat{\theta}'$. Therefore, by proposition 4.2, $\hat{\theta}'$ cannot be bunched. Consequently, a non-local incentive constraint has to bind to $\hat{\theta}'$. If $\hat{\theta}'$ is isolated in the set of types to which non-local incentive constraints are binding, $\eta(\theta)$ has to be discontinuous at $\hat{\theta}'$ by the definition of $\hat{\theta}'$. Then also $q(\theta)$ is discontinuous at $\hat{\theta}'$. But this is impossible by lemma 4.3. Hence, $\hat{\theta}'$ cannot be isolated in the set of types to which non-local incentive constraints bind.

It remains to show the closedness part of the lemma. Note first that a monotone solution is continuous almost everywhere. Consequently, the principal's payoff is not changed if $q(\cdot)$ is changed at its discontinuity points. I want to resolve this ambiguity using the following convention: Say $q(\theta)$ is discontinuous at θ'. Then $q(\theta') = q^-(\theta')$ if there exists an increasing sequence of types θ_i $i = 1, 2, \ldots$ such that (i) $\lim_{i\to\infty} \theta_i = \theta'$ and (ii) a non-local incentive constraint is binding from or to each θ_i. If such a sequence does not exist, $q(\theta') = q^+(\theta')$.

With this convention in mind, consider a sequence of types θ_n with $n = 1, 2, \ldots$ such that a non-local incentive constraint is binding from each θ_n to some $\hat{\theta}_n$. Assume that $\lim_{n\to\infty} \theta_n = \theta'$. Then it has to be shown that $\Phi(\theta', \hat{\theta}') = 0$ for some $\hat{\theta}'$. Since all $\hat{\theta}_n$ belong to the closed and bounded interval $[\underline{\theta}, \bar{\theta}]$, there is a convergent subsequence of $\hat{\theta}_n$.

[95] Note that θ' cannot be bunched because of proposition 4.4 and $q^-(\theta') = q^r(\theta')$.

I will denote the elements of this subsequence by $\hat{\theta}_k$ with $k = 1, 2, \ldots$. The corresponding type from which a non-local incentive constraint is binding to $\hat{\theta}_k$ is denoted by θ_k. Now, take $\hat{\theta}' = \lim_{k\to\infty} \hat{\theta}_k$. Note that there always exists a *monotone* subsequence of θ_k. It is therefore without loss of generality to assume θ_k to be monotone. For concreteness, assume $\theta_{k+1} \geq \theta_k$ for all $k = 1, 2, \ldots$. As $\Phi(\theta_k, \hat{\theta}_k) = 0$ for all $k = 1, 2, \ldots$, continuity of $\Phi(\cdot)$ at $(\theta', \hat{\theta}')$ is sufficient for $\Phi(\theta', \hat{\theta}') = 0$. As $\pi(\cdot)$ is continuous by local incentive compatibility and $c(\cdot)$ is continuous by assumption, continuity of $\Phi(\cdot)$ at $(\theta', \hat{\theta}')$ follows if $q(\cdot)$ is continuous at $\hat{\theta}'$. Since θ_k is monotonically increasing, continuity from below is actually sufficient. But this is ensured by the convention above.

If $\theta_{k+1} \leq \theta_k$ for all $k = 1, 2, \ldots$, the convention establishes $q(\hat{\theta}') = q^+(\hat{\theta}')$ which is needed in this case.

The proof for the closedness of the set of types to which non-local incentive constraints bind works in the same way. *Q.E.D.*

Proof of lemma 4.5: From lemma 4.3, non-local incentive constraints cannot bind from any $\theta \in [\theta_s^b, \theta_e^b)$. To satisfy similar properties as in theorem 4.1, $\eta(\theta)$ has therefore to be non-decreasing on (θ_s^b, θ_e^b).

Let $\eta(\theta)$ be defined by (4.5) for all types that are not bunched. Define $\eta(\theta)$ on the bunching interval using the following two step procedure: First, all $\hat{\theta} \in (\theta_s^b, \theta_e^b)$ such that $\Phi(\theta, \hat{\theta}) = 0$ and (C1') as well as (C2') are satisfied are assigned $\eta(\hat{\theta}) = \eta(\theta)$. Second, types in $\theta \in (\theta_s^b, \theta_e^b)$ who are not assigned a value for $\eta(\theta)$ in step 1 are assigned the same η as the highest type $\theta' < \theta$ that was already assigned a value $\eta(\theta')$.

Now it is shown that the constructed $\eta(\theta)$ is non-decreasing on (θ_s^b, θ_e^b): Say, there are two types $\hat{\theta}_1, \hat{\theta}_2 \in (\theta_s^b, \theta_e^b)$ with $\hat{\theta}_2 > \hat{\theta}_1$ which are assigned an η in the first step. Then (C2') implies that $\theta_1 > \theta_2$. From theorem 4.1 and the structure of the solution as depicted in figure 4.5, it follows that $\eta(\theta_2) \geq \eta(\theta_1)$. Therefore, $\eta(\hat{\theta}_2) \geq \eta(\hat{\theta}_1)$. The second step does not change the monotonicity of $\eta(\theta)$ which proves that $\eta(\theta)$ is non-decreasing on (θ_s^b, θ_e^b).

If non-local incentive constraints are not binding for the bunched types, no type is assigned a value for $\eta(\theta)$ in step 1. Consequently, $\eta(\theta)$ is constant on (θ_s^b, θ_e^b).

Next, it is shown that $\eta(\theta)$ is also non-decreasing at the types θ_s^b and θ_e^b. First, note that the proof of theorem 4.1 can be easily extended to show that $\eta(\theta_s^b) \leq \eta(\theta_e^b)$: If this inequality did not hold, reduce $q(\theta)$ on $(\theta_s^b - \varepsilon, \theta_s^b)$ and increase $q(\theta)$ marginally on $(\theta_e^b, \theta_e^b +$

ε) such that $\int_{\theta_s^b-\varepsilon}^{\theta_e^b+\varepsilon}\int_{q(\theta_s^b-\varepsilon)}^{q(t)} c_{q\theta}(q,t)\,dq\,dt$ remains the same before and after the change. As in the proof of theorem 4.1, this change would increase the principal's payoff without impeding incentive compatibility (note that non-local incentive constraints cannot bind from the bunched types because of lemma 4.3). Consequently, $\eta(\theta_s^b) \leq \eta(\theta_e^b)$.

Second, it is necessary to show that–with the above constructed $\eta(\theta)$ on (θ_s^b, θ_e^b)–there is no upward jump of $\eta(\theta)$ at θ_e^b (no downward jump of $\eta(\theta)$ at θ_s^b). If no type is assigned an η in the first step of the procedure above, this is obvious. Therefore, take the case where some type in the bunching interval is assigned a value $\eta(\theta)$ in the first step of the procedure. Then the claim follows from theorem 4.1: Say, $\eta^-(\theta_e^b) = \eta(\theta_1)$ for some type θ_1 from which a non-local incentive constraint binds. The structure of the solution (as depicted in figure 4.5) and theorem 4.1 imply that $\eta^+(\theta_e^b) = \eta^-(\theta_1)$.[96] Since $\eta(\theta)$ is non-increasing at θ_1 according to theorem 4.1, it follows that $\eta^-(\theta_1) \geq \eta^+(\theta_1)$ and therefore $\eta^-(\theta_e^b) \geq \eta^+(\theta_e^b)$. A similar argument holds for θ_s^b.

It remains to show $\int_{\theta_s^b}^{\theta_e^b} \nu_\theta(\theta)\,d\theta = 0$. But this follows directly from $\nu(\theta_s^b) = \nu(\theta_e^b) = 0$. *Q.E.D.*

Proof of proposition 4.3: By lemma 4.3, $q(\theta)$ cannot be discontinuous at a type to which a non-local incentive constraint binds (with the exception of boundary types of bunching intervals). Therefore, theorem 4.1 implies that a solution could only be discontinuous at types where $\eta(\theta)$ is non-increasing or at the boundary types of a bunching interval to which a non-local incentive constraint is binding.

First, it is shown that $\eta(\theta)$ is also non-increasing at such boundary types of a bunching interval. To see this take a bunching interval $[\hat{\theta}_1, \hat{\theta}_2]$ to which non-local incentive constraints bind and suppose the solution was discontinuous at $\hat{\theta}$, i.e. $q^-(\hat{\theta}_2) < q^+(\hat{\theta}_2)$. By the arguments in the proof of lemma 4.3, $\int_{\hat{\theta}_2}^{\theta} c_{q\theta}(q^-(\hat{\theta}_2), t)\,dt > 0$ for any θ such that $\Phi(\theta, \hat{\theta}_2) = 0$. But then an argument as in the proof of theorem 4.1 applies: There is an incentive compatible way to increase $q(\hat{\theta})$ for $\hat{\theta} \in [\hat{\theta}_2 - \varepsilon, \hat{\theta}_2]$ and decrease the decision for types in $[\hat{\theta}_2, \hat{\theta} + \varepsilon]$. Incentive compatible means that binding non-local incentive constraints are not violated and the decision remains monotone (details in the proof of

[96] If non-local incentive constraints bind from types $\theta' \in (\theta_e^b, \theta_1)$ to types $\hat{\theta}' \in (\theta_e^b, \theta_1)$, this holds still true because of the necessary condition (C3). Also discontinuities at $\theta'' \in (\theta_e^b, \theta_1)$ do not matter as by lemma 4.3 and theorem 4.1 $\eta(\theta)$ is non-increasing at θ''. If there are several bunching intervals, the argument holds for the highest interval and given this, it holds for the second highest etc..

theorem 4.1). If $\eta(\cdot)$ was strictly increasing at $\hat{\theta}_2$, such a change would increase the principal's payoff. Therefore, $\eta(\cdot)$ has to be decreasing at $\hat{\theta}_2$. A similar argument applies at $\hat{\theta}_1$. A discontinuity is only possible at $\hat{\theta}_1$ if $\int_{\hat{\theta}_1}^{\theta} c_{q\theta}(q(\hat{\theta}_1), t)\, dt < 0$ for all θ such that $\Phi(\theta, \hat{\theta}_1) = 0$. Therefore, decreasing the decision on $[\hat{\theta}_1, \hat{\theta}_1 + \varepsilon]$ and increasing the decision on $[\hat{\theta}_1 - \varepsilon, \hat{\theta}_1)$ can be done in an incentive compatible way. If $\eta(\cdot)$ was strictly increasing, such a change would increase the principal's payoff.

Hence, $q(\theta)$ can only be discontinuous at types where $\eta(\theta)$ is non-increasing. Second, it is shown that a discontinuity in $q(\theta)$ would lead to an upward jump of $\eta(\theta)$ at the discontinuity type which implies that there cannot be a discontinuity in $q(\theta)$.

By local incentive compatibility, $q(\theta)$ can only jump upwards, i.e. $q^-(\theta') < q^+(\theta')$ at a hypothetical discontinuity type θ'. Using the definition of $\eta(\theta)$ in (4.5) one can calculate the change in $\eta(\theta')$ at the discontinuity type

$$\begin{aligned}\eta^+(\theta') - \eta^-(\theta') &= \int_{q^-(\theta')}^{q^+(\theta')} \frac{d\,\eta(\theta')}{d\,q(\theta')}\, dq \\ &= \int_{q^-(\theta')}^{q^+(\theta')} \frac{(u_{qq} - c_{qq}) f c_{q\theta} + (1-F) c_{qq\theta} c_{q\theta} - (u_q - c_q) f c_{qq\theta} - (1-F) c_{q\theta} c_{qq\theta}}{c_{q\theta}^2}\, dq\end{aligned}$$

where all functions are evaluated at (q, θ'). Note that the integrand is positive whenever $q \leq q^{fb}(\theta')$. If $q > q^{fb}(\theta')$, the integrand can be written as

$$\frac{f(u_q - c_q)}{c_{q\theta}} \left(\frac{u_{qq} - c_{qq}}{u_q - c_q} - \frac{c_{qq\theta}}{c_{q\theta}} \right)$$

which is also positive due to the condition of the proposition. Hence, $\eta(\theta)$ would jump up at θ' but this contradicts that $q(\theta)$ can only be discontinuous at types where $\eta(\theta)$ is non-increasing. *Q.E.D.*

Proof of proposition 4.4: The proof is by contradiction. Suppose the optimal decision $q(\theta)$ was above the first best decision for some types. Since there is no distortion at the top by assumption and since the optimal decision cannot drop discontinuously downward (local incentive compatibility), there has to be a type θ' at which the optimal decision intersects $q^{fb}(\theta)$ from above. The proof works now in two steps. First, I show that a non local incentive constraint must bind from θ' and second that then non local incentive compatibility is violated for some type close to θ'.

Note that $q(\theta) > q^{fb}(\theta)$ if and only if $\eta(\theta) > 1 - F(\theta)$. Since $1 - F(\theta)$ is decreasing and $q(\theta) > (<) q^{fb}(\theta)$ slightly above (below) θ', it follows that $\eta_\theta(\theta')$ is negative. But

then, by theorem 4.1, a non local incentive constraint has to be binding from θ' to some $\hat{\theta}'$. Furthermore, the necessary condition $\int_{q(\hat{\theta}')}^{q(\theta')} c_{q\theta}(q,\theta')\,dq = 0$ has to hold.

Next consider a type $\theta'' = \theta' - \epsilon$ with $\epsilon > 0$ very small. Since $q^m(\theta)$ is increasing and $\int_{q(\hat{\theta})}^{q(\theta)} c_{q\theta}(q,\theta')\,dq = 0$, clearly $\int_{q(\hat{\theta}')}^{q^{fb}(\theta'')} c_{q\theta}(q,\theta'')\,dq < 0$. Since $q(\theta'') > q^{fb}(\theta'')$, it has to hold that $\int_{q(\hat{\theta}')}^{q(\theta'')} c_{q\theta}(q,\theta'')\,dq < 0$ as well. The same inequality holds for all $\theta \in (\theta'', \theta')$. But then $\Phi(\theta'', \hat{\theta}') = \Phi(\theta', \hat{\theta}') + \int_{\theta''}^{\theta'} \int_{q(\hat{\theta}')}^{q(t)} c_{q\theta}(q,t)\,dq\,dt < 0$, i.e. incentive compatibility from θ'' to $\hat{\theta}'$ is violated. Hence, the optimal decision cannot be above the first best decision.

Continuity of the optimal decision is now straightforward: $q(\theta) \leq q^{fb}(\theta)$ implies that $1 - F(\theta) - \eta(\theta) \geq 0$. Therefore, the left hand side of the first order condition $u_q - c_q + (1 - F - \eta)c_{q\theta} = 0$ is strictly decreasing in q. The same arguments as in the proof of proposition 4.3 show that $q(\theta)$ has to be continuous.

Last it has to be shown that the decision is strictly monotone when it is below first best and $u_{q\theta} \geq 0$. This will be done in two steps. The first step is to show that $q(\theta)$ is strictly increasing if $\eta_\theta(\theta) \geq 0$. The second step is to show that in a hypothetical bunching interval there are types θ at which $\eta_\theta(\theta) \geq 0$ which by the first step contradicts that these types are bunched.

First, the decision $q(\theta)$ has to satisfy

$$[u_q(q(\theta),\theta) - c_q(q(\theta),\theta)] + \frac{(1 - F(\theta) - \eta(\theta))}{f(\theta)} c_{q\theta}(q(\theta),\theta) = 0 \tag{4.9}$$

by theorem 4.1. From the implicit function theorem, the sign of $q_\theta(\theta)$ can be determined. Note that $q(\theta) \leq q^{fb}(\theta)$ implies $1 - F(\theta) - \eta(\theta) \geq 0$. This in turn implies that the derivative of the left hand side of (4.9) with respect to q is negative. Hence, the sign of $q_\theta(\theta)$ is the sign of the partial derivative of the equation above with respect to θ. Denoting $(1 - F(\theta) - \eta(\theta))$ by $\lambda(\theta)$ this derivative is

$$u_{q\theta}(q(\theta),\theta) - c_{q\theta}(q(\theta),\theta) + \frac{\lambda(\theta)}{f(\theta)} c_{q\theta\theta} + \frac{\partial\, \lambda(\theta)/f(\theta)}{\partial\theta} c_{q\theta}(q(\theta),\theta). \tag{4.10}$$

Now take a bunching interval $[\theta_1, \theta_2]$ (closed or open). The first three terms are clearly positive as $q(\theta_1) \leq q^{fb}(\theta_1)$ implies $\lambda(\theta) \geq 0$. The fourth term is positive if $\eta_\theta(\theta) \geq 0$ as then

$$\frac{\partial\, \lambda(\theta)/f(\theta)}{\partial\theta} = \frac{-f^2(\theta) - f_\theta(\theta)(1 - F(\theta))}{f^2(\theta)} - \frac{\eta_\theta(\theta)}{f(\theta)} + \frac{f_\theta(\theta)\eta(\theta)}{f^2(\theta)} < 0$$

where the inequality comes from the monotone hazard rate assumption if $f_\theta(\theta) \leq 0$. If $f_\theta(\theta) > 0$, then $q^{fb}(\theta) \geq q(\theta)$ implies $\lambda(\theta) \geq 0$ which ensures the inequality above.

Now turn to the second step. Suppose contrary to the proposition that an interval (θ_1, θ_2) exists in which types are bunched and non-local incentive constraints are either binding to these types or are lax.[97] Using the same argument as in the proof of theorem 4.1, it becomes evident that $\eta(\theta)$ as defined by (4.5) cannot be decreasing on the whole interval (θ_1, θ_2): If this was the case, increasing $q(\theta)$ for types $((\theta_2 + \theta_1)/2, \theta_2)$ and decreasing $q(\theta)$ slightly for the other bunched types would increase the principal's payoff (and can be done in an incentive compatible way). From the definition of $\eta(\theta)$ and the differentiability of q on the bunching interval, it follows that $\eta(\theta)$ is continuous and differentiable on this interval. Consequently, there has to be some type in the interior of the bunching interval where $\eta_\theta(\theta) \geq 0$. But then the first step shows that this type cannot be bunched. *Q.E.D.*

Proof of proposition 4.5: Take two types θ' and $\hat{\theta}'$ such that a non-local incentive constraint is binding from θ to $\hat{\theta}$ under the optimal decision $q(\theta)$. By (C3), $\eta(\theta') = \eta(\hat{\theta}')$ and for this proof η (in $\Phi^\eta()$) simply denotes this common value $\eta(\theta') = \eta(\hat{\theta}')$.

First, suppose that $(\theta', \hat{\theta}')$ does not minimize $\Phi^\eta(\theta, \hat{\theta})$ on $[\hat{\theta}', \theta]$ and call the minimizer $(\theta'', \hat{\theta}'')$. Then incentive compatibility under the optimal decision requires $\Phi(\theta'', \hat{\theta}'') \geq 0$. If $q(\theta)$ was $\tilde{q}(\theta)$ for all types in $[\hat{\theta}', \hat{\theta}''] \cup [\theta'', \theta']$, then $\Phi(\theta', \hat{\theta}') = \Phi^\eta(\theta', \hat{\theta}') + \Phi(\theta'', \hat{\theta}'') - \Phi^\eta(\theta'', \hat{\theta}'') > 0$ where the inequality stems from the definition of $(\theta'', \hat{\theta}'')$ as global minimizer of $\Phi^\eta(\theta, \hat{\theta})$. Therefore ic would not be binding between θ' and $\hat{\theta}'$.

If $q(\theta) \neq \tilde{q}(\theta)$ for some types in $[\hat{\theta}', \hat{\theta}''] \cup [\theta'', \theta']$, then ic must be binding for some of these types.[98] But this will only relax ic, i.e. $q(\theta) > \tilde{q}(\theta)$ in a monotone solution. Therefore $\Phi(\theta', \hat{\theta}')$ will be even higher than when $q(\theta) = \tilde{q}(\theta)$ and therefore ic cannot bind between θ' and $\hat{\theta}'$. This is the desired contradiction. Consequently, $(\theta', \hat{\theta}')$ has to minimize $\Phi^\eta(\theta, \hat{\theta})$ on $[\hat{\theta}', \theta']$.

Second, suppose that $(\theta'', \hat{\theta}'')$ with $\hat{\theta}'' < \hat{\theta}' < \theta' < \theta''$ has $\Phi^\eta(\theta', \hat{\theta}') > \Phi'(\theta'', \hat{\theta}'')$. In fact choose θ'' and $\hat{\theta}''$ such that it is the global minimizer of $\Phi^\eta(\theta, \hat{\theta})$ under the constraint $\hat{\theta} < \hat{\theta}' < \theta' < \theta$.

Now suppose for the moment that all types in $[\hat{\theta}'', \hat{\theta}'] \cup [\theta', \theta'']$ had $q(\theta) = \tilde{q}(\theta)$. Then since $\Phi(\theta', \hat{\theta}') = 0$ but $(\theta'', \hat{\theta}'')$ minimizes $\Phi^\eta(\theta, \hat{\theta})$, ic would be violated for θ'' and $\hat{\theta}''$.

[97] By lemma 4.3, types from which non-local incentive constraints bind cannot be bunched.

[98] Because of lemma 4.2 ic cannot bind from outside $[\hat{\theta}', \theta]$ into the interval (neither the other way round).

If $q(\theta) \neq \tilde{q}(\theta)$ for some types in $[\hat{\theta}'', \hat{\theta}'] \cup [\theta', \theta'']$, then ic was binding for some types in those intervals. In a monotone solution this implies that $q(\theta) < \tilde{q}(\theta)$ for these types. Put differently, ic is stricter under $q(\theta)$ than under $\tilde{q}(\theta)$.[99] But then ic will be even more violated for θ'' and $\hat{\theta}''$ under $q(\theta)$ than under $\tilde{q}(\theta)$. Therefore, there cannot be a global minimizer $(\theta'', \hat{\theta}'')$ with $\hat{\theta}'' < \hat{\theta}' < \theta' < \theta''$. *Q.E.D.*

Proof of corollary 4.1: Note first that the highest type θ from which a non-local incentive constraint is binding must have $q(\theta) = q^r(\theta)$ if θ is interior. This follows from the reasoning in the proof of lemma 4.4. The same holds for the lowest type $\hat{\theta}$ to which a non-local incentive constraint binds. Therefore, there is a type pair such that (i) $q(\theta') = q^r(\theta')$, (ii) $q(\hat{\theta}') = q^r(\hat{\theta}')$ and (iii) $\Phi(\theta', \hat{\theta}') = 0$.

Since $(\theta', \hat{\theta}')$ satisfy (C2) and (C1) with q^r and given the results of proposition 4.5, $(\theta', \hat{\theta}')$ locally minimize $\Phi^r(\theta, \hat{\theta})$. Proposition 4.5 rules out that $\hat{\theta}^r < \hat{\theta}' < \theta' < \theta^r$ and also $\hat{\theta}' < \hat{\theta}^r < \theta^r < \theta'$. Hence, it still has to be shown that there cannot be an overlap between the two type pairs, i.e. $\hat{\theta}' < \hat{\theta}^r < \theta' < \theta^r$ or $\hat{\theta}^r < \hat{\theta}' < \theta^r < \theta'$. To get a contradiction suppose $\hat{\theta}' < \hat{\theta}^r < \theta' < \theta^r$. In a similar way as in lemma 4.2, one can now show that in this case $\Phi^r(\theta^r, \hat{\theta}') < \Phi^r(\theta^r, \hat{\theta}^r)$ thereby contradicting that $(\theta^r, \hat{\theta}^r)$ is the global minimizer of $\Phi^r(\theta, \hat{\theta})$:

$$\begin{aligned}\Phi^r(\theta^r, \hat{\theta}') &= \Phi^r(\theta^r, \hat{\theta}^r) + \Phi^r(\theta', \hat{\theta}') + \int_{\hat{\theta}^r}^{\theta'} \int_{q^r(\hat{\theta}^r)}^{q^r(t)} c_{q\theta}(q,t)\, dq\, dt - \int_{\theta'}^{\theta^r} \int_{q^r(\hat{\theta}')}^{q^r(\hat{\theta}^r)} c_{q\theta}(q,t)\, dq\, dt \\ &= \Phi^r(\theta^r, \hat{\theta}^r) + \Phi^r(\theta', \hat{\theta}') - \Phi^r(\theta', \hat{\theta}^r) - \int_{\theta'}^{\theta^r} \int_{q^r(\hat{\theta}')}^{q^r(\hat{\theta}^r)} c_{q\theta}(q,t)\, dq\, dt\end{aligned}$$

By proposition 4.5, $\Phi^r(\theta', \hat{\theta}') - \Phi^r(\theta', \hat{\theta}^r) \leq 0$. Furthermore, $\int_{q^r(\hat{\theta}')}^{q^r(\theta')} c_{q\theta}\, dq = 0$ since $(\theta', \hat{\theta}')$ locally minimize $\Phi^r(\theta, \hat{\theta})$. Therefore, $\int_{q^r(\hat{\theta}')}^{q(\hat{\theta}^r)} c_{q\theta}\, dq > 0$ as $q^r(\theta') > q^r(\hat{\theta}^r)$ and $c_{qq\theta} < 0$. From $c_{q\theta\theta} > 0$ it follows that $\int_{\theta'}^{\theta^r} \int_{q^r(\hat{\theta}')}^{q^r(\hat{\theta}^r)} c_{q\theta}(q,t)\, dq\, dt > 0$ which shows that $\Phi^r(\theta^r, \hat{\theta}') < \Phi^r(\theta^r, \hat{\theta}^r)$. This is the desired contradiction.

A similar argument can be made for the case $\hat{\theta}^r < \hat{\theta}' < \theta^r < \theta'$. Consequently, the only possibility is that $(\theta', \hat{\theta}') = (\theta^r, \hat{\theta}^r)$ which had to be shown.

[99] Strictly speaking one also has to show that ic did not bind from outside $[\hat{\theta}'', \theta'']$ into this interval (or the other way round), thereby increasing $q(\theta)$ for some types in say (θ', θ''). If however this was the case and the increase in $q(\theta)$ was such that ic between θ'' and $\hat{\theta}''$ was relaxed by it, then there has to exist a type $\hat{\theta}''' \in (\theta', \theta'')$ and a type $\theta''' > \theta''$ with $\Phi(\theta''', \hat{\theta}''') = 0$ and $q(\hat{\theta}''') = \tilde{q}(\hat{\theta}''')$. But this would contradict that $(\theta'', \hat{\theta}'')$ is a global minimum of $\Phi^\eta(\theta, \hat{\theta})$ (analogously to the proof of lemma 4.2), i.e. $\Phi'(\theta''', \hat{\theta}'') < \Phi'(\theta'', \hat{\theta}'')$.

If the highest/lowest type from/to which a non-local incentive constraint is binding is a boundary type, this type's decision is not necessarily the relaxed decision. However, the minimization argument does not change which concludes the proof. *Q.E.D.*

Proof of proposition 4.6: First, note that under the conditions of proposition 4.1 one can focus on decisions above $s(\theta)$: If some $q(\theta) < s(\theta)$ was used in a stochastic contract with positive probability, the principal could do better by assigning $q^s(q(\theta), \theta)$ instead of $q(\theta)$. The proof is equivalent to the one of proposition 4.1.[100]

Second, suppose the optimal contract was stochastic and denote by $G(q, \theta)$ the distribution of q at type θ. Consider now an alternative deterministic contract $q^*(\theta)$ where $q^*(\theta) \geq s(\theta)$ is determined such that $c_\theta(q^*(\theta), \theta) = \int_q c_\theta(q, \theta)\, dG(q, \theta)$. In short, the slope of the rent function $\pi(\theta)$ and therefore the rent of each type remains the same under both contracts.[101] It will be shown that under the assumptions of proposition 4.6 this change increases the principal's payoff and relaxes incentive compatibility.

Since only $q(\theta) \geq s(\theta)$ have to be considered, there is a one to one relationship between q and $-c_\theta(q, \theta)$. Define $h(z, \theta) \geq s(\theta)$ as the decision corresponding to $-c_\theta$ being z, i.e. $z = -c_\theta(h(z, \theta), \theta)$. Then the principal's objective can be written as

$$W = \int_{\underline{\theta}}^{\bar{\theta}} [u(h(z, \theta), \theta) - c(h(z, \theta), \theta)] f(\theta) - [1 - F(\theta)] z \, d\theta. \tag{4.11}$$

The next step is to show that W is concave in z. This implies that the deterministic decision q^* increases the principal's payoff. The last step will then be to show that this deterministic decision is also incentive compatible.

Using $h_z(z, \theta) = \frac{1}{-c_{q\theta}(h, \theta)}$, which follows from the definition of $h(z, \theta)$, it is straightforward to derive

$$\frac{\partial^2 W}{\partial z^2} = \int_{\underline{\theta}}^{\bar{\theta}} \frac{c_{qq\theta}(h, \theta)}{c_{q\theta}^2(h, \theta)} \left[\frac{u_{qq}(h, \theta) - c_{qq}(h, \theta)}{c_{qq\theta}} - \frac{u_q(h, \theta) - c_q(h, \theta)}{c_{q\theta}(h, \theta)} \right] f(\theta)\, d\theta.$$

By condition (CVR), the integrand is negative and therefore W is concave in z.

[100] Admittedly, it is not obvious whether lemma 4.1 holds for stochastic contracts. Therefore, I show here explicitly that upward incentive constraints are relaxed if all $q(\theta) < s(\theta)$ are substituted by $q^s(q(\theta), \theta)$. Take $\hat{\theta} > \theta$, then the incentive constraint can be written as $\int_\theta^{\hat{\theta}} \int_q c_\theta(q, t)\, dG(q, t) - \int_q c_\theta(q, t)\, dG(q, \hat{\theta})\, dt \geq 0$. Changing $q(\theta) < s(\theta)$ to $q^s(q(\theta), \theta)$ does not change the first term. But since $\int_{q(\theta)}^{q^s(q(\theta), \theta)} c_{q\theta}(q, t)\, dq < 0$ for all $t < \hat{\theta}$, the change relaxes the incentive constraint through the second term.

[101] It is straightforward to check that local incentive compatibility requires the slope of the rent function under the stochastic contract to be $\int_q c_\theta(\tilde{q}, \theta)\, dG(q, \theta)$.

Incentive compatibility of q^* means that for arbitrary types θ and $\hat{\theta}$

$$\Phi^*(\theta,\hat{\theta}) \equiv \int_{\hat{\theta}}^{\theta} c_\theta(q^*(\hat{\theta}),t) - c_\theta(q^*(t),t)\,dt \geq 0.$$

To verify this, it is useful to see that (4.6) implies that $-c_{\theta\theta}(h(z,\theta),\theta)$ is convex in z:

$$\frac{d^2\{-c_{\theta\theta}(h(z,\theta),\theta)\}}{dz^2} = \frac{-1}{c_{q\theta}}\frac{\partial \frac{c_{q\theta\theta}}{c_{q\theta}}}{\partial q} \geq 0$$

This convexity implies

$$\frac{d}{d\theta}\left\{\int_q -c_\theta(q,\theta)\,dG(q,\hat{\theta}) + c_\theta(q^*(\hat{\theta}),\theta)\right\} = \int_q -c_{\theta\theta}(q,\theta)\,dG(q,\hat{\theta}) + c_{\theta\theta}(q^*(\hat{\theta}),\theta) \geq 0.$$

The last inequality implies that $\int_{\hat{\theta}}^{\theta}\int_q -c_\theta(q,t)\,dG(q,\hat{\theta}) + c_\theta(q^*(\hat{\theta}),t)\,dt$ is a convex function of θ with a minimum at $\theta=\hat{\theta}$ where the function value is 0. Consequently,

$$\int_{\hat{\theta}}^{\theta} c_\theta(q^*(\hat{\theta}),t)\,dt \geq \int_{\hat{\theta}}^{\theta}\int_q c_\theta(q,t)\,dG(q,\hat{\theta})\,dt.$$

But then,

$$\Phi^*(\theta,\hat{\theta}) = \int_{\hat{\theta}}^{\theta} c_\theta(q^*(\hat{\theta}),t) - c_\theta(q^*(t),t)\,dt \geq \int_{\hat{\theta}}^{\theta}\int_q c_\theta(q,t)\,dG(q,\hat{\theta})\,dt - c_\theta(q^*(t),t)\,dt \geq 0$$

where the last inequality follows from the incentive compatibility of the stochastic contract $G(q,\theta)$ and the definition of q^*. *Q.E.D.*

4.8.3. Existence of an optimal contract

This appendix shows that an optimal contract exists and therefore the characterization done in the paper is meaningful. It is assumed that $q^v(q,\theta) \geq q^s(q,\theta)$ for all $q \in [0, q^f(\theta)]$ and all $\theta \in [\underline{\theta},\bar{\theta}]$ and therefore proposition 4.1 applies. Before showing existence, two useful lemmata are derived.

Define $\tilde{q}$ such that $\int_0^{\tilde{q}} c_{q\theta}(q,\bar{\theta})\,dq = 0$. Since $c_{qq\theta} < 0$, $\tilde{q}$ is unique and therefore properly defined.

Lemma 4.6. *Any incentive compatible contract with a decision $q(\theta)$ above $\bar{q} = max\{q^{fb}(\bar{\theta}),\tilde{q}\}$ for some type is dominated by a contract consisting of decision*

$$q^c(\theta) = min\{q(\theta),\bar{q}\}$$

and transfers such that $\pi(\underline{\theta}) = 0$ and $\pi_\theta(\theta) = \int_{\underline{\theta}}^{\theta} -c_\theta(q(t),t)\,dt$.

Proof. The concavity of the virtual valuation implies that the principal's payoff under $q^c(\theta)$ is higher than under $q(\theta)$. Hence, the lemma holds if the changed contract is incentive compatible.

Note that incentive compatibility of $q^c(\theta)$ is obvious if $q(\theta) > \bar{q}$ for all θ. Now define $\theta^m = inf\{\theta : \ q(\theta) > \bar{q}\}$. Note that incentive compatibility from θ^m to any lower type is not affected by the change from $q(\cdot)$ to $q^c(\cdot)$ since $\Phi(\theta^m, \hat{\theta})$ does not change.

The next step is to see that $q(\theta) > \bar{q}$ for all $\theta > \theta^m$. The reason is that local incentive compatibility does not allow for any decision in $[s(\theta), \bar{q}]$ as long as $q(\theta)$ stays above $s(\theta)$. Furthermore, downward jumps to a decision below $s(\theta)$ would require that $\int_{q^+(\theta^j)}^{q^-(\theta^j)} c_{q\theta}(q, \theta^j)\, dq \geq 0$ at the jump type θ^j (for local incentive compatibility). But by the definition of $\bar{q}$ and from $c_{q\theta\theta} > 0$, this inequality cannot hold for any type below $\bar{\theta}$ (and a jump at the boundary type $\bar{\theta}$ would not hurt the following argument).

Therefore, all types above θ^m will have $\bar{q}$ as their changed decision. From lemma 4.1 it follows that only incentive compatibility from types above θ^m to types below θ^m has to be checked. Therefore take an arbitrary $\theta > \theta^m$ and some $\hat{\theta} < \theta^m$. Then $\Phi(\theta, \hat{\theta}) = \Phi(\theta^m, \hat{\theta}) - \int_{\theta^m}^{\theta} \int_{q(\hat{\theta})}^{\bar{q}} c_{q\theta}(q, t)\, dq\, dt > 0$ where the inequality follows from the incentive compatibility between θ^m and $\hat{\theta}$ under $q(\theta)$, the definition of $\tilde{q}$ and $c_{q\theta\theta} > 0$.

Q.E.D.

Lemma 4.7. *Take a sequence of incentive compatible decision functions*[102] $q^n(\theta) \leq \bar{q}$, $n = 1, 2 \ldots$, *and let this sequence converge to* $q(\theta)$. *Then* $q(\theta)$ *is incentive compatible.*

Proof. Define $\tilde{c}_{q\theta} = \max_{q \in [0,\bar{q}],\ \theta \in [\underline{\theta}, \bar{\theta}]} |c_{q\theta}(q, \theta)|$. Since $[0, \bar{q}] \times [\underline{\theta}, \bar{\theta}]$ is compact and $c_{q\theta}(\cdot)$ is continuous by assumption, $\tilde{c}_{q\theta}$ exists.

Now suppose contrary to the lemma that $\Phi(\theta, \hat{\theta}) = -\varepsilon$ for some θ, $\hat{\theta} \in \Theta$ and $\varepsilon > 0$ and therefore incentive compatibility is violated under $q(\theta)$. From convergence of $\{q^n(\theta)\}$, for each $\delta > 0$ there exists an N_δ such that $|q^n(\theta) - q(\theta)| \leq \delta$ for all types and all $n > N_\delta$. Therefore,

$$\Phi(\theta, \hat{\theta}) = \int_{\hat{\theta}}^{\theta} \int_{q(\hat{\theta})}^{q(t)} -c_{q\theta}(q, t)\ dq\, dt \geq \int_{\hat{\theta}}^{\theta} \int_{q^n(\hat{\theta})}^{q^n(t)} -c_{q\theta}(q, t)\ dq\, dt - \int_{\hat{\theta}}^{\theta} 2\delta \tilde{c}_{q\theta}\ dt$$

for an arbitrary $n > N_\delta$. But then choosing a $\delta < \frac{\varepsilon}{2\tilde{c}_{q\theta}(\theta - \underline{\theta})}$ shows that $\Phi(\theta, \hat{\theta}) > -\varepsilon$ as

[102] An incentive compatible decision is a decision such that the menu consisting of this decision and transfers defined by $\pi(\theta) = \int_{\underline{\theta}}^{\theta} -c_\theta(q(t), t)\, dt$ is incentive compatible.

$\Phi^n(\theta, \hat{\theta}) \geq 0$ where $\Phi^n(\cdot)$ denotes $\Phi(\cdot)$ under $q^n(\cdot)$. This contradicts the definition of ε and therefore $q(\theta)$ is incentive compatible. *Q.E.D.*

Given proposition 4.1 and the previous two results, the existence proof in Jullien (2000) applies. For completeness, I replicate the proof briefly. The problem of the principal is the program:

$$\max_{q(\theta)} \int_{\underline{\theta}}^{\bar{\theta}} (u(q(\theta), \theta) - c(q(\theta), \theta)) f(\theta) + (1 - F(\theta)) c_\theta(q(\theta), \theta) \, d\theta$$

subject to

$$\begin{aligned} \Phi(\theta, \hat{\theta}) &\geq 0 \quad \text{for all} \quad \theta, \hat{\theta} \in [\underline{\theta}, \bar{\theta}] \\ 0 &\leq q(\theta) \leq \bar{q} \end{aligned}$$

Let W^* be the maximum value of the program. Take a sequence of decision functions such that $q^n(\theta)$ induce a value larger than $W^* - \frac{1}{n}$ and each $q^n(\theta)$ is incentive compatible . Because of proposition 4.1, the sequence can be chosen such that each $q^n(\theta)$ is an increasing function. Then Helly's selection theorem, see Billingsley (1986) Thm. 25.9, yields that there exists a non-decreasing function $q(\theta)$ which is the limit of a subsequence $q^{n_k}(\theta)$ at every point of continuity of $q(\theta)$ and therefore almost everywhere on $[\underline{\theta}, \bar{\theta}]$. Lebesgue's dominated convergence theorem, see Billingsley (1986) Thm. 16.4, yields that the principal's payoff under $q(\theta)$ is W^*. By lemma 4.7, $q(\theta)$ is implementable and therefore an optimal contract exists.

Chapter 5

Cost incentives for doctors: A double-edged sword

5.1. Introduction

It is well known that insurance creates moral hazard: In the health sector, insured people would like to have more expensive treatments than socially optimal. On the other hand, treatments are normally prescribed by doctors. If doctors took the costs of treatment into account in their treatment decision, the moral hazard problem should disappear. The tradition in the medical profession, however, is to view oneself as advocate of one's patients. Consequently, the patient's wellbeing is put first and costs are only secondary. What is more, doctors are often explicitly hostile towards cost incentives in doctor remuneration. The German chamber of doctors, for instance, writes in its principles of health policy[103]

> [...] the role of the doctor as advocate for his patient must not be restricted [...] The state must not establish financial schemes (e.g. bonus-malus system) which could suggest to the patient that materialistic, self-serving aspects are also of importance for medical decisions.

It is important to understand whether the doctors' concerns are mainly self-interested, e.g. worries about reputation and pay, or whether financial incentives for doctors could have a negative impact on social welfare. Put differently, can patient advocacy be interpreted as an efficient institutional response to the particular structure of the health care market? Answering this question will also give some insight into the optimal design of health care markets. In particular, in which parts of the health care system should cost

[103] Translation by the author. Original title and source: "Gesundheitspolitische Leitsätze der Ärzteschaft–Ulmer Papier" Beschluss des deutschen Ärztetags 2008, Anlage 1, p. 6, http://www.bundesaerztekammer.de/downloads/UlmerPapierDAET111.pdf

incentives for doctors be employed and where are cost incentives less likely to succeed?

This paper focuses on the communication between patient and doctor. The patient's input, e.g. describing his symptoms and their intensity, is vital to reach the right diagnosis.[104] The main mechanism I explore in this paper is the following: Patients are (fully) insured. If doctors take costs into account in their treatment decision, their objectives and the objectives of their patients are no longer aligned.[105] Such a misalignment undermines the patient's trust in his doctor which in turn affects communication negatively.[106] More technically, in a setting where the patient has private information, e.g. about his symptoms and their intensity, he has the possibility to exaggerate his symptoms (or their intensity) in order to get a more expensive treatment. Of course, the doctor will anticipate such strategic exaggerating. This anticipation gives the patient further incentives to exaggerate and so on. The appropriate model to analyze such a "rat race" is the cheap talk framework. This paper will therefore extend the canonical cheap talk model to the imperfect information setting typical for the health sector. Although a complete breakdown of communication can be prevented, communication will be worse in equilibrium because of the misalignment of interests, i.e. less information is transmitted from patient to doctor. It is shown that this communication effect can make a system without cost incentives preferable from a social welfare point of view. If the patient's collaboration is hardly needed, a system with cost incentives is preferable. For example, a doctor can easily establish that a patient has a broken leg by having an X-ray. The symptoms reported by the patient are less important in this case. If, on the other hand, an illness might have a psychological background, the patient's collaboration is essential and a system without cost incentives might be preferable.

[104] The importance of communication is also stressed in the aforementioned document of the German chamber of doctors where it is stated that "health can neither be commanded nor produced since health depends crucially on the patient's collaboration." Also there is a whole string of the medical literature dealing with doctor-patient communication, see Stewart (1995) for a survey.

[105] Negative effects from cost incentives on the doctor-patient relationship are also established in the medical literature, see for example Rodwin (1995), Kao et al. (1998) or Gallagher and Levinson (2004).

[106] There is no doubt that patients understand this nexus: According to Gallagher et al. (2001) 73% of their respondents dislike the idea of a cost control bonus for their doctor and 91% favor disclosure to the patient if such a bonus was in place. Furthermore, 95% of those who dislike the bonus stated that the bonus would lower their trust in their physician.

The main idea of the paper is a tradeoff between having the best information to base the decision on and having the socially best decision rule. A related tradeoff is known in organization theory. Alonso et al. (2008) ask how much autonomy division managers should have. Giving division managers more autonomy results in better information use in decision making but less coordination across divisions. In my paper, the only way to get better information (from the patient) is a socially less desirable decision rule for the doctor. In both papers, there is a tradeoff between the quality of information and the quality of the decision rule (for a given information structure). The downside of an informed decision in Alonso et al. (2008) is a lack of coordination while in my paper it will be the neglect of costs. A major technical difference between the papers is that division managers (headquarters manager) have full (no) information in Alonso et al. (2008) while doctor and patient will both receive a noisy signal in my model. This setup seems to be closer to reality in the health sector.

From a technical point of view, the paper contributes to the cheap talk literature following the seminal paper by Crawford and Sobel (1982). Their model is extended in de Barreda (2010) to a setup where the decision maker receives a noisy signal. My paper generalizes further by substituting the perfect information on the sender/expert/patient side by a noisy signal.

This paper complements existing literature on the design of health care systems. Early contributions as Arrow (1963) and Pauly (1968) already point out the moral hazard caused by health insurance: Insured patients might overconsume treatment from a social welfare perspective because they are insured. Ma and McGuire (1997) introduce the physician as an additional player and analyze contractual difficulties in the health market. In particular, health outcome and doctor's effort are non-contractible and even the quantity of care consumed can be subject to misreporting. Ma and McGuire (1997) analyze how these contractual constraints influence optimal contracts between insurance and patient as well as between insurance and physician. My paper focuses on a different kind of constraint, i.e. a constraint in information transmission arising in the communication between doctor and patient. It will be shown that the necessity of information transmission between patient and doctor might constrain the power of the incentive scheme offered to the doctor.

Obviously related is the literature on physician compensation and managed care. In his survey of the managed care literature, Glied (2000) mentions two problems of "supply-side cost sharing," i.e. cost incentives for physicians: (i) underprovision of necessary services and (ii) strong incentives to avoid costly cases. In this context, my paper adds a third problem: Hampered information transmission between doctor and patient. Furthermore, my paper provides one possible explanation for the ambiguous cost effect of managed care mentioned in Glied (2000).

The medical literature contains statements like "payment arrangements could significantly undermine patients' beliefs that their physicians are acting as their agents" (Mechanic and Schlesinger, 1996) and emphasizes that there should be no conflict of interest between patient and doctor (Emanuel and Dubler, 1995). Kao et al. (1998) find that patients trust their physician less if the physician is capitated than when he is payed on a fee for service basis. Physicians are also less satisfied with their relationships with capitated patients compared to their average patient (Kerr et al., 1997). My paper contributes by formalizing why trust, interpreted as shared objectives, is vital for the patient-physician relationship. Such a formalization is interesting for two reasons: First, it allows for both costs (less trust) and benefits (less overtreatment) of cost incentives. Second, one can obtain results concerning the optimal design of health care systems, i.e. where in the health system are aligned interests especially important and where could cost incentives improve welfare.

The next section introduces the model and is followed by a simple numerical example. This example illustrates the main points. Section 5.4 analyzes a general model and answers the question: When do cost incentives work? The final section concludes by discussing certain assumptions and pointing out testable predictions as well as possible applications in different areas.

5.2. Formal setting

Patient and doctor have a common prior F over the set of all possible health states of the patient. The set of health states is denoted by Θ. A health state can be interpreted either as the severity of a given disease or as a set containing different diseases. The

patient receives a private signal $\sigma^p \in \Sigma^p$ about his health state. In practice this signal can be interpreted as the symptoms a patient can report to his doctor or as the intensity of his symptoms. The doctor receives also a private signal $\sigma^d \in \Sigma^d$ about the patient's health. This signal can be interpreted as the result of the doctor's examination, e.g. his interpretation of an X-ray photograph or listening to the patient's heartbeat. Given the health state, there is a distribution $G(\sigma^p, \sigma^d|\theta)$ of signals which is common knowledge. Put differently, $G(\sigma^p, \sigma^d|\theta)$ gives the probabilities that a patient (doctor) receives signal σ^p (σ^d) given a health state θ.

The timing is the following: First, the patient's health state is determined by nature. This health state is unknown to doctor and patient. Second, doctor and patient receive their signals $\sigma = (\sigma^p, \sigma^d)$ which correspond to the true health state through G. Third, the patient can send a message, e.g. communicating his signal, to the doctor. Fourth, the doctor determines a treatment τ from a set of available treatments. The costs of the treatment $c(\tau)$ are paid for by the patient's insurance.

Utility of the patient depends only on his true health state $\theta \in \Theta$ and the treatment τ. In particular, a patient's well being does in the end not depend on the signals. For the doctor, I look at two scenarios: Either the doctor has "no cost incentives" which means that he makes his treatment decision to maximize the patient's utility or he is "cost sensitive" (or "has cost incentives") with which I mean that he maximizes social welfare. Social welfare is the patient's utility minus costs. The perspective of the paper is therefore eventually the perspective of a (benevolent) designer of the health system, e.g. a government or an insurance plan, who has to determine which kind of incentives he gives to the doctor.

5.3. A simple example

This section deals with a small numerical example which illustrates that cost incentives can lead to lower welfare. Take $\Theta = \{A, B, C\}$ and $\Sigma^p = \Sigma^d = \{0, 1\}$. In words, there are three diseases called A, B and C. One can interpret health states either as similar diseases or as levels of severity of the same disease. Doctor and patient will each receive one of two possible signals which are denoted by 0 and 1. For example, the patient's

signal could be whether he feels "no/little pain" or "strong pain" while the doctor's signal could be whether the patient's heartbeat is unusual or not. The prior F is given by disease A and B occurring with probability 2/5 each and disease C with probability 1/5. The distribution G is given in the following table:

prior	2/5	2/5	1/5
σ	A	B	C
(0,0)	4/5	0	0
(0,1)	0	0	1
(1,0)	0	4/5	0
(1,1)	1/5	1/5	0

The interpretation is that, given health state A, signal $(\sigma^p, \sigma^d) = (0, 0)$ occurs with probability 4/5 and signal $(\sigma^p, \sigma^d) = (1, 1)$ occurs with probability 1/5. Assume that there are three available treatments which are denoted by a, b and c. The patient's utility and the costs of each treatment are given in the following table:

	A	B	C	costs
a	8	5	8	5
b	10	10	9	8
c	1	1	10	0

To illustrate: A patient with disease A receiving treatment a has a utility of 8. Treatment a costs 5. Therefore, welfare would be $8 - 5 = 3$ in this situation.

One interpretation is that "disease" C is being healthy and treatment c is the option "no treatment". Treatment b is a very effective but also very expensive treatment while a is not so effective but substantially cheaper. A quick calculation shows that treatment a is welfare maximizing in health state A where welfare is defined by patient utility minus costs. The same is true for b in health state B and c in C.

5.3.1. No cost incentives

If the doctor has no cost incentives, the incentives of doctor and patient are aligned. The patient will therefore communicate his true signal σ^p in equilibrium.[107] The doctor can

[107] In principle, there is also a pooling equilibrium in which the doctor takes only his own signal into account and the patient sends the same message regardless of his signal. However, this equilibrium is Pareto dominated and does not seem very realistic.

then base his decision on both signals and maximizes gross consumer surplus. Hence, the doctor knows the disease whenever the signals are $(0,0)$, $(0,1)$ or $(1,0)$. If the signal is $(1,1)$, the doctor assigns equal probabilities to disease A and B. This leads to the following optimal decisions: $(0,0) \to b$, $(0,1) \to c$, $(1,0) \to b$ and $(1,1) \to b$

Expected welfare is therefore:[108]

$$W^{nci} = \frac{16}{50}(10-8) + \frac{1}{5}10 + \frac{16}{50}(10-8) + \frac{4}{50}(10-8) + \frac{4}{50}(10-8) = \frac{180}{50} \qquad (5.1)$$

5.3.2. Cost sensitive doctor

If the doctor is cost sensitive, his preferred decisions (if he knew both signals) would be: $(0,0) \to a$, $(0,1) \to c$, $(1,0) \to b$ and $(1,1) \to b$. Hence, there is a conflict between the patient and the doctor whenever the signal is $(0,0)$: The doctor prefers treatment a while the patient prefers b. Next, I write down the optimal decision of the doctor if he only knows his own signal σ^d. If $\sigma^d = 0$, he assigns equal probability to disease A and B. Therefore, the optimal treatment is b. If $\sigma^d = 1$, he assigns probability $2/9$ to disease A, $2/9$ to disease B and $5/9$ to disease C. It is straightforward to calculate that in this case the optimal treatment is c.

In principle, there could be two kinds of equilibrium: First, a separating equilibrium in which the patient truthfully reports his signal to the doctor, i.e. the two signals are separated. Second, a pooling equilibrium in which the patient sends the same message regardless of his signal.

Suppose there is a separating equilibrium, i.e. the patient communicates his signal σ^p truthfully to the doctor in equilibrium. The doctor will then implement the welfare maximizing treatment knowing both signals. If $\sigma^p = 0$, the patient expects–given his signal–to get a utility of $u^{truth} = 8/13 * 8 + 5/13 * 10 = 114/13$.[109] If however the patient lied and communicated $\sigma^p = 1$, the doctor would implement treatment b and the patient's expected utility would be $u^{lie} = 8/13 * 10 + 5/13 * 9 = 125/13$. Hence, lying pays off for the agent and there cannot be a separating equilibrium.

[108] Just to illustrate: The first term is the probability of being in state A and receiving the signal $(0,0)$, i.e. $2/5 * 4/5$, multiplied with the utility of the resulting treatment b in state A, i.e. 10, minus the costs of this treatment, i.e. 8.

[109] Given $\sigma^p = 0$, the patient assigns probability $8/13$ to health state A with signal $\sigma = (0,0)$ which leads to treatment a. With the counter probability $5/13$, he expects state C with signal $\sigma = (0,1)$ and treatment c.

Consequently, there is a pooling equilibrium in which the doctor uses only his own signal. Welfare is then

$$W^c = \frac{16}{50}(10-8) + \frac{1}{5}10 + \frac{16}{50}(10-8) + \frac{4}{50}1 + \frac{4}{50}1 = \frac{172}{50} \tag{5.2}$$

Since $W^c < W^{nci}$, cost incentives reduce welfare in this example. Nevertheless, costs are lower if the doctor is cost sensitive since the signal $(1,1)$ leads to the low cost treatment c while b is prescribed without cost incentives. The driving force behind this result are the conflicting objectives of patient and doctor which result in a break down of communication.

5.3.3. Variation I: Restricting the choice set

Interestingly, there is an easy fix in this example: Suppose, the health authority does not clear treatment a. Hence, treatment a is not available. But then there is no conflict between doctor and patient as even a cost sensitive doctor will now prescribe b if the signal $(0,0)$ occurs. Unfortunately, this means that cost incentives simply do not matter/work: Every signal leads to the same treatment with and without cost incentives. Furthermore, this trick will not always work: Amend the example above with a disease D which can be identified with certainty (so there would be a signal $(2,2)$ which occurs if and only if the health state is D). If in this state D treatment a is by far superior to all other treatments, a health authority banning treatment a would reduce welfare.

5.3.4. Variation II: Increasing costs

The negative information effect of cost incentives can be so strong that costs can be higher under cost incentives. To see this, change the example above by changing the ex ante probability of disease C from $1/5$ to $p^c < 1/5$ and assign the ex ante probability $(1-p^c)/2$ to sickness A and B. Note that this does not change decisions without cost incentives as it is always perfectly known whether one is in state C or not.

If, however, p^c is small enough and the doctor knows only his own signal, he will prescribe treatment b instead of treatment c when he receives signal $\sigma^d = 1$. This inevitably leads to higher costs than without cost incentives: Now b is always prescribed while c was prescribed without cost incentives for signal $\sigma = (0,1)$. Note that a lower p^c will make the incentive constraint of a separating equilibrium even tougher, i.e. reducing

p^c does not lead to a separating equilibrium. It turns out that in the example expected costs are higher with cost incentives if $p^c < 2/41$.

This result is slightly reminiscent of the empirical results concerning the cost effects of managed care. One feature of many managed care plans are cost incentives for doctors, e.g. cost control boni or capitation payment. As Glied (2000) reports in his survey, results on the cost effect of managed care are however inconclusive: Some studies report higher costs, some report lower costs or no cost difference between managed care and traditional care plans.

5.4. Model and results

This section uses a more general model to analyze the setting and effect described before. There are two reasons why this is desirable: First, one has to verify that the effects described above are not due to the discrete nature of the example. Second, this will allow to determine under which circumstances cost incentives are welfare maximizing and therefore have implications for the optimal design of a health care system.

The patient's message in the example above is "cheap talk": The message itself does not have direct payoff implications. Only the treatment decision is relevant for the patient's utility and welfare. The canonical model for cheap talk games is Crawford and Sobel (1982). To fit the health sector, the information structure of Crawford and Sobel (1982) has to be amended as described below.

I assume that health state θ is a real number from some bounded interval and also σ^p, σ^d and τ are assumed to be real numbers.[110] Without loss of generality take $\Theta = [0, 1]$. Again one can interpret the health state either as the severity of a given disease or one views Θ as a continuum of diseases. Higher signals are assumed to imply higher expected states. To make this formal define by $H(\theta|\sigma^p, \sigma^d)$ the cumulative distribution function which gives the probability that the state is below θ given signals σ^p and σ^d. This distribution is derived from $F(\theta)$ and $G(\sigma^p, \sigma^d|\theta)$ using Bayes' rule. The assumption is that $H(\theta|\sigma^p, \sigma^d)$ first order stochastically dominates $H(\theta|\sigma^{p\prime}, \sigma^{d\prime})$ whenever $\sigma^d \geq \sigma^{d\prime}$ and $\sigma^p \geq \sigma^{p\prime}$. In words, a higher signal implies that higher health states are more likely to

[110] Restricting τ to some interval, e.g. $\mathbb{R}_+$ is possible as explained in footnote 114. Drawing the signals from some closed subset of $\mathbb{R}$ simplifies matters, see assumption 5.1.

occur.

Patient utility $u(\theta-\tau)$ is a function of "distance" between health state and treatment. It is assumed that the patient is fully insured, i.e. costs of treatment do not enter his utility function. Assume that $u(\theta-\tau)$ is two times continuously differentiable, strictly concave and attains its maximum at 0. Put differently, patient utility is maximized if $\tau=\theta$ and is lower the further away treatment τ is from this ideal treatment. A treatment above (below) θ corresponds to overtreatment (undertreatment) from the patient's point of view. It is *not* assumed that $u(\cdot)$ is symmetric and therefore over- and undertreatment might affect utility in different ways. The cost function $c(\tau)$ is strictly increasing and marginal costs are bounded away from 0, i.e. $c'(\tau)\geq\delta \quad \forall\tau$ for some $\delta>0$. This last assumption implies that the patient's utility is never aligned with the social objective or, put differently, the patient always prefers a more expensive treatment than socially optimal because he is insured. If there was no such conflict, cost incentives would simply not matter for the outcome. Consequently, introducing cost incentives could not even help to reduce costs.

The solution concept is Perfect Bayesian Nash Equilibrium. After observing his signal σ^p a patient updates his beliefs about his health state θ and about the doctor's signal. Given σ^p, a strategy for the patient is a probability distribution over Σ^p denoted by $q(m|\sigma^p)$.[111] This distribution gives the probability of reporting $m\in\Sigma^p$ when the true signal is σ^p. For illustration purposes, think of a partition equilibrium in which patients with signals in, say, $[0.3,0.4]$ are bunched, i.e. send the same message. In this case $q(m|\sigma^p)$ could be a uniform distribution over $[0.3,0.4]$ for all $\sigma^p\in[0.3,0.4]$. Given his signal σ^d and the message he receives from the patient, the doctor updates his beliefs about the health state of the patient θ and chooses his preferred treatment. For simplicity, I assume that $u(\theta-\tau)-c(\tau)$ is strictly concave in τ which implies that there is a unique socially efficient treatment τ^w. This assumption is, for example, satisfied if $c(\tau)$ is linear or convex. Hence, the doctor will always have a unique preferred treatment which I denote by $\tau^d(m,\sigma^d)$. The strategies $(q(m|\sigma^p),\tau^d(m,\sigma^d))$ form an equilibrium if:

1. For each σ^p, $q(m|\sigma^p)$ is a distribution, i.e. $\int_0^1 q(m|\sigma^p)\,dm=1$, and if $q(m^*|\sigma^p)>0$ then $m^*\in argmax_m\int_0^1\int_{\Sigma^d}u(|\theta-\tau^d(m,\sigma^d)|)\,dP(\theta,\sigma^d|\sigma^p)$ where $P(\theta,\sigma^d|\sigma^p)$ is the

[111] For notational convenience $q(m|\sigma^p)$ is a probability density function but mass points can be easily accommodated.

distribution of (θ, σ^d) derived from $G(\sigma^p, \sigma^d|\theta)$ and $F(\theta)$ conditional on observing σ^p and using Bayes' rule.[112]

2. For each m and σ^d, treatment maximizes the doctor's objective. For the cost sensitive doctor, this means that $\tau^d(m, \sigma^d) = argmax_\tau \int_0^1 [u(\theta - \tau) - c(\tau)]\, dH(\theta|m, \sigma^d)$ where with a slight abuse of notation $H(\theta|m, \sigma^d)$ is the distribution of the health state conditional on observing σ^d and m using by Bayes' rule (given $G(\sigma^p, \sigma^d|\theta)$, $F(\theta)$ and $q(m|\sigma^p)$). Without cost incentives $\tau^d(m, \sigma^d) = argmax_\tau \int_0^1 u(\theta - \tau)\, dH(\theta|m, \sigma^d)$.

In words, the first condition says that the patient reports with positive probability only signals maximizing his utility given the strategy of the doctor. The second condition establishes that the doctor uses an optimal strategy given the patient's equilibrium behavior.

For the analysis of this model the following technical assumption proves to be useful. Note that the boundedness part is automatically satisfied if H_{σ^p} is continuous and the signal σ is drawn from a closed set, i.e. if Σ^p and Σ^d are closed intervals.

Assumption 5.1. *$H(\theta|\sigma^p, \sigma^d)$ is differentiable in σ^p and $|H_{\sigma^p}(\theta|\sigma^p, \sigma^d)|$ is bounded from above by some $M > 0$. At all states where $H(\theta|\sigma^p, \sigma^d)$ has a density $h(\theta|\sigma^p, \sigma^d)$, this density is also differentiable in σ^p and h_{σ^p} is bounded.*

Put differently, beliefs about the true health state do not change too sharply if the patient's signal changes marginally. Note that slightly irregular distribution, e.g. with mass points at a "healthy state" $\theta = 0$, can be allowed. Assumption 5.1 simplifies the analysis by ensuring that the doctor's treatment decision is differentiable in the patient's signal. In fact, it implies that there is an upper bound on how strongly the doctor's treatment decision reacts to a marginal change in σ^p (in a hypothetical situation in which the doctor knows the patient's signal). Loosely speaking, this means that a patient who exaggerates his signal a little bit will–as a consequence–only get a slightly higher treatment. See the proof of proposition 5.1 for details.

The game is then similar to the information transmission model of Crawford and Sobel (1982) with three additional twists: First, the doctor (receiver in the language of

[112] Note that the patient takes expectations not only over the health state but also over the doctor's signal because σ^d will influence the doctor's treatment decision.

Crawford and Sobel) receives a signal while he is completely ignorant in Crawford and Sobel (1982). Second, the patient (sender) does not know the state of the world. Instead, he has a noisy signal. Third, the divergence of interests between doctor (receiver) and patient (sender) is not fixed but depends on the treatment (decision). The following proposition extends results from Crawford and Sobel (1982) to this more general setting.

Proposition 5.1. *With cost incentives, there exists no separating equilibrium. There exist partitioning equilibria on the range of σ^p. Each part of this partition has a minimum length κ which is bounded away from zero. If Σ^p is bounded, the number of parts in the partition is bounded from above.*

Proof. see appendix

The intuition is the following: In equilibrium, a patient cannot tell his true signal to the doctor. If he did, the doctor would prescribe a treatment that is "too cheap" from the patient's point of view (as the patient does not care about costs). Hence, the patient would have an incentive to overstate his signal. In practice, this would mean to claim additional symptoms or to overstate the intensity of existing symptoms. What happens in equilibrium is that the patient's signal range is partitioned and the patient reports in which part of the partition his signal lies. The doctor does not know the precise signal of the patient but gets a rough idea which he takes into consideration when choosing the treatment. Because of the partitioning, a patient can no longer overstate his signal "a little bit". If the patient deviated by reporting a higher part of the partition, he would get a substantially higher treatment. In equilibrium he will not deviate because he expects this treatment to be too high. One could interpret this in the following two ways: First, a patient does not want to report symptoms that are too much different from the real ones as this could mislead the doctor, i.e. result in treating the wrong illness. Second, extreme overstatement of symptoms could result in too strong medication with severe side effects. Hence, the patient does not want to overstate his existing symptoms too much.

It is also clear that the partition cannot be arbitrarily fine: If the parts are too small, then overstating one's signal "a little bit" is again possible. This explains the minimum length statement in the proposition. The minimum part length immediately implies that the number of parts is bounded if the interval from which patient signals are drawn is bounded.

The mechanism through which cost incentives can harm welfare is the same as in the example of section 5.3: If the objectives of doctor and patient are different, the patient has an incentive to use his information strategically to get the more expensive treatment he wants. In equilibrium, the doctor will have less information (partitioning of signal range) compared to the situation without cost incentives. Consequently, he is more prone to make inappropriate treatment decisions. In short, there are two effects when introducing cost incentives: First, costs are taken into account which, ceteris paribus, decreases costs and increases welfare. Put differently, the doctor stops prescribing excessively expensive treatments. Second, communication and therefore the information of the doctor is worse. Hence, treatment decisions are less accurate which reduces welfare. Whether the cost or the information effect dominates is ex ante unclear. The following propositions show that in two extreme cases the cost effect dominates and therefore cost incentives lead to higher welfare than no cost incentives.

Proposition 5.2. *Welfare is higher with cost incentives if the doctor's signal is sufficiently informative. That is, given $G(\sigma^p, \sigma^d|\theta)$, there exists an $\varepsilon > 0$ such that cost incentives lead to higher welfare than no cost incentives if the doctor's signal is drawn from $\varepsilon G(\sigma^p, \sigma^d|\theta) + (1-\varepsilon)1_\theta$ where 1_θ is a distribution putting all probability mass on θ. Cost incentives lead also to higher welfare if the patient's signal is sufficiently uninformative, i.e. for $\varepsilon > 0$ small enough if the patient's signal is drawn from $\varepsilon G(\sigma^p, \sigma^d|\theta) + (1-\varepsilon)U_\theta$ where U_θ is the uniform distribution over $[0, 1]$.*

Proof. see appendix

This result is intuitive: If the doctor is able to determine the patient's health state almost on his own, i.e. without knowing the patient's signal, then the patient's signal is useless. Therefore, the information effect of introducing cost incentives is small while the cost effect is still there.

One interpretation of proposition 5.2 is that cost incentives become eventually more attractive with medical progress. This holds at least true if medical progress implies better diagnosis possibilities for doctors. Consequently, one might then expect to see more cost incentive elements in health care systems over time.

A second interpretation is that some specialists optimally should have cost incentives while others should not. A radiologist or a trauma surgeon will normally base his deci-

sions on his own examination and less on the patient's report. This might be less true for an internist or a general practitioner.

A related third interpretation is that an optimal health care system should incorporate selective cost incentives. More precisely, cost incentives should be applied for the treatment of diseases where the doctor's information is relatively more important than the patient's information.

Proposition 5.3. *Cost incentives lead to higher welfare than no cost incentives if social and private objectives differ sufficiently. That is, for any given information structure and cost function $c(\tau)$ there exists an $\alpha > 0$ such that cost incentives lead to higher welfare than no cost incentives under the cost function $\alpha c(\tau)$.*

Proof. see appendix

The intuition is that the cost effect will become dominant if (marginal) costs are high enough. Consequently, the information loss due to cost incentives is negligible compared to the cost effect.

In line with previous interpretations cost incentives are especially useful for specialists dealing with high cost treatments on a regular basis. Also diseases involving high cost treatment on a regular basis are especially well suited for cost incentives.

The previous propositions illustrate when cost incentives are superior to no cost incentives. Now, I want to give an example where no cost incentives are superior to cost incentives. In fact, I can use the same example as Crawford and Sobel (1982) which is attractive for two reasons: First, it is very simple and allows therefore for an analytical solution. Second, it has been used repeatedly in the cheap talk literature and has become a benchmark example there.

Example 5.1. *Health states are uniformly distributed on $[0,1]$. The patient has perfect knowledge of the health state while the doctor's signal is completely uninformative. Assume that the patient's utility function is a quadratic loss function, i.e. $u(\theta,\tau) = -(\theta-\tau)^2$, and that the cost function is linear in treatment, i.e. $c(\tau) = \alpha\tau$. Given the information that σ^p (which is now the true health state) is in the interval (s_1, s_2), the optimal treatment decision for a doctor with cost incentives is $\tau = \frac{s_1+s_2-\alpha}{2}$. With $\alpha = 1/10$ the model is mathematically equivalent to the example in Crawford and Sobel (1982). It is shown there that the finest possible equilibrium partition is $(0, 2/15, 7/15, 1)$, i.e.*

a patient will report whether his signal is in $[0, 2/15)$ *or in* $[2/15, 7/15)$ *or in* $[7/15, 1]$. *Utility of a patient with state* θ *in* $[0, 2/15)$ *is given by* $-(1/60-\theta)^2$, *with* $\theta \in (2/15, 7/15)$ *utility is* $-(1/4-\theta)^2$ *and with* $\theta \in (7/15, 1)$ *utility is* $-(41/60-\theta)^2$. *Expected consumer utility in this partition equilibrium is therefore*

$$EU = \int_0^{2/15} -\left(\frac{1}{60}-\theta\right)^2 d\theta + \int_{2/15}^{7/15} -\left(\frac{1}{4}-\theta\right)^2 d\theta + \int_{7/15}^{1} -\left(\frac{41}{60}-\theta\right)^2 d\theta \approx -0.01058$$

while expected costs are

$$EC = \frac{1}{10}\left(\frac{2}{15}\frac{1}{60} + \frac{5}{15}\frac{15}{60} + \frac{8}{15}\frac{41}{60}\right) = 0.045.$$

Hence expected welfare is $-0.01058 - 0.045 = -0.05558$. *Note that this is an upper bound on welfare: Of course, there are also equilibria with partitions consisting of only two parts or one part. It is easy to check that these equilibria result in lower welfare.*

Without cost incentives the patient will truthfully reveal his signal and therefore communicate the true health state to the doctor. Consequently, $\tau = \theta$ *and consumer welfare is* 0. *Expected costs are* $\frac{1}{10}0.5 = 0.05$ *which results in expected welfare of* -0.05. *Therefore, no cost incentives lead to higher welfare than cost incentives.*

To conclude this section, think of a generalization: Say, the planner could set to which extent the doctor takes costs into account, i.e. the doctor maximizes the expected value of $u(\theta - \tau) - \beta c(\tau)$ with his treatment decision where $\beta \in [0, 1]$ is set by the planner. I want to argue that $\beta = 1$ is normally not optimal. Hence, a welfare maximizing planner does not want a welfare maximizing doctor. The idea is the following: Say $\beta = 1$ induces the partition $[s_0, s_1, \ldots, s_N]$. Now suppose β is decreased marginally (starting from $\beta = 1$) and assume for now that the partition remained the same: The doctor will prescribe higher treatments in response. As he was welfare maximizer before this change, this will only have a second order effect on welfare (if the partition did not change). However, there will be a first order information effect: Because the doctor prescribes higher treatments, a patient with signal $\sigma^p = s_1$ is no longer indifferent but strictly prefers to report in the (s_0, s_1) part of the partition. Consequently, a patient with $\sigma^p = s_1$ will only be indifferent if s_2 is reduced. Hence, the partition can become finer.

Of course, there are some caveats to this intuition: For example, it might be hard in practice to fine tune doctor's incentives in such a precise way. In particular, doctors

will be heterogenous in how much they respond to incentives. Nevertheless, the intuition above suggests that a welfare maximizing planner does not want doctors to naively maximize welfare also in more general settings than the one discussed in this paper.

5.5. Discussion and conclusion

Introducing cost incentives for doctors turns out to be a double-edged sword: On the one hand, taking costs into consideration should avoid the prescription of too expensive treatments. On the other hand, misalignment of patient's and doctor's incentives will hamper communication between the two: The patient has an incentive to exaggerate and in equilibrium this leads to signal bunching. Consequently, the doctor has worse information and is less likely to assess the patient's health state correctly. Knowing about the uncertainty he might even choose more expensive treatments to be on the safe side. In a numerical example, this can lead to higher costs than under no cost incentives (see section 5.3).

If costs are very high or if the doctor is able to assess the health state very accurately given only his signal, cost incentives are the welfare maximizing policy. This shows that an optimal health care system will use different degrees of cost incentives in different circumstances. In practice, cost incentives could differ across diseases and across specialists.

One could interpret trust as congruence in decision making, e.g. shared objectives. In this interpretation, the model in this paper allows to formalize the idea that trust is important in the patient-doctor relationship. A lack of trust reduces the quality of communication and eventually the quality of the doctor's diagnosis. This effect could constrain contracting between insurances and doctors.

Note that some seemingly strong assumptions are actually not very restrictive: The concentration on two extreme cases where the doctor either maximizes patient utility or total welfare is obviously not realistic. The main effect, that diverging objectives lead to worse communication, however, holds true whenever the doctor cares more about costs than the patient. By the same argument, it is not restrictive to assume full indemnity insurance: The main point is that the patient does not bear the full social costs which

is a feature of any form of insurance. The results do therefore not depend on a specific form of insurance. One can interpret the costs $c(\tau)$ simply as the part of treatment costs paid by the patient's health insurance.

In some sense, the model is a best case scenario for the benevolent designer: He can freely set the doctor's incentives without incurring any costs. In practice setting up an incentive scheme for doctors might actually be costly. Doctors might also not respond immediately because of previously formed habits. It is therefore even more remarkable that the designer might not want to give cost incentives to the doctor in the model of this paper.

The model gives several testable predictions. Quality of diagnosis should decrease after an introduction of cost incentives for doctors. Such a quality decrease could be reflected in the data in different ways: First, therapies could be changed more often (if the doctor realizes the error at a later stage). Second, patients with a given diagnosis-treatment pair will be treated less successfully (e.g. take longer to recover) because some receive the wrong treatment due to a wrong diagnosis. These effects should be more pronounced for specialists and diseases where patient input is vital for the diagnosis. If trust reflects the willingness to communicate, one should expect patient's trust in their doctor to be lower when their doctor has cost incentives. This last result is indeed confirmed by the health literature, see for example Kao et al. (1998).

More abstract, a welfare maximizing sponsor (say a benevolent government) might prefer a decision maker (doctor) who shares his preferences not with the sponsor but with the patient. In a broader context an agent might benefit from surrendering his interests when information provision by another party is important. This could have applications in other contexts like mediation: A mediator with decision power who shares the interests of another party might be preferable to making the decision oneself.

In general, shared objectives prove to be vital for information provision. Patient advocacy can therefore be seen as an institutional response to the importance of information provision by patients. Consequently, one might expect similar institutions to emerge whenever information provision by affected parties is vital. In this context, the relationship between a lawyer and his client could serve as an additional example.

5.6. Appendix: Proofs

Proof of proposition 5.1: The proof proceeds in a number of steps. The first three steps establish that there cannot be a separating equilibrium, i.e. there is no equilibrium in which a patient always reports his true signal. Consequently, patients with some signals are bunched together. Patients in one "bunch" (one part of a partition of the signal range) send the same report to the doctor. Steps four and five establish that each part of a partition must have a minimum length, i.e. the partition cannot be arbitrarily fine.

The first step is to show that there exists a $b > 0$ such that $argmax_\tau \int_0^1 [u(\theta - \tau) - c(\tau)]\, dH(\theta|m, \sigma^d) + b \leq argmax_\tau \int_0^1 u(\theta - \tau)\, dH(\theta|m, \sigma^d)$ for a given equilibrium strategy $q(m|\sigma^p)$; i.e. the patient would opt for an at least b higher treatment than a cost sensitive doctor if he chose (and had the same information). This follows from the first order conditions corresponding to the two $argmax$ expressions

$$\int_0^1 -u'(\theta - \tau)\, dH(\theta|m, \sigma^d) = \begin{cases} c'(\tau) \\ 0 \end{cases} . \tag{5.3}$$

The left hand side of (5.3) is continuous in τ and also strictly decreasing in τ. Since $c'(\tau) \geq \delta > 0$ and $u'(\cdot)$ is continuous, the claim follows. This argument is for a given (m, σ^d) but the infimum of all these b over (m, σ^d) will also be strictly positive. To establish this, it is sufficient to show that the derivative of the left hand side of (5.3) with respect to τ is bounded:[113] Since $u'(\theta - x) > 0$ for $x \geq 1$ and any $\theta \in [0, 1]$, the optimal treatment is bounded from above by 1. Furthermore, the optimal treatment is bounded from below by $\underline{\tau}$ solving $u'(-\underline{\tau}) = c'(\underline{\tau})$, i.e. the optimal treatment if the doctor knew that $\theta = 0$. Therefore $-1 \leq \theta - \tau \leq 1 - \underline{\tau}$. By the continuity of $u''(\cdot)$ and the compactness of $[-1, 1 - \underline{\tau}]$, $u''(\cdot)$ is bounded on this interval. Consequently, the derivative of the left hand side of (5.3) is a weighted (by the distribution $H(\cdot)$) average of a bounded function and therefore bounded. Denote by $B > 0$ such a bound on the derivative of the left hand side of (5.3). Then we can choose $b = \delta/B$.[114]

[113] Just to illustrate why boundedness is sufficient: Say the derivative of the left hand side of (5.3) is between 0 and $-B$. Since this left hand side is differentiable, the two τ solving (5.3) with the right hand side equal to zero and equal to $c'(\tau)$ have to differ by at least δ/B.

[114] If the treatment is restricted to be larger than, say, 0, the argument still holds true as long as

Second, the patient's expected utility is under separating higher under a slightly higher decision than the cost sensitive doctor takes. From the first step and the strict concavity of $u(\cdot)$ it follows that any treatment in $(\tau^d, \tau^d + b)$ yields a higher expected utility for the patient than τ^d.

Third, in a hypothetical separating equilibrium the patient attains a higher utility by misrepresenting slightly upwards as the doctor will increase his decision uniformly continuously in σ^p. The implicit function theorem gives for a hypothetical separating equilibrium

$$\frac{d\tau^d}{d\sigma^p} = \frac{\frac{\partial \int_0^1 -u'(\theta-\tau)\, dH(\theta|\sigma^p,\sigma^d)}{\partial \sigma^p}}{-\int_0^1 [u''(\theta-\tau) - c''(\tau)]\, dH(\theta|\sigma^p,\sigma^d)}. \tag{5.4}$$

The denominator is obviously positive as it is (-1) times the second order condition of the doctor's maximization problem. The numerator is positive as well because of stochastic dominance: As $-u'(\theta-\tau)$ is a strictly increasing function of θ, we have $\int_0^1 -u'(\theta-\tau)\, dH_1(\theta) > \int_0^1 -u'(\theta-\tau)\, dH_2(\theta)$ whenever $H_1(\theta)$ first order stochastically dominates $H_2(\theta)$. Since $H(\theta|\sigma^{p\prime}, \sigma^d)$ first order stochastically dominates $H(\theta|\sigma^p, \sigma^d)$ whenever $\sigma^{p\prime} > \sigma^p$, the numerator has to be positive. The uniform continuity follows from the boundedness of 5.4: The numerator is bounded by assumption 5.1 and the fact that $u'(\theta-\tau)$ is bounded on the relevant range. The strict concavity of the doctor's program implies that the denominator is strictly bounded away from zero.[115] By uniform continuity, misrepresentation can be chosen small enough to prevent an "overreaction" by the doctor.

Consequently, there cannot be a separating equilibrium. The same argument shows that also locally, i.e. on some subinterval of the patient's signal range, there cannot be a perfect separation of types, i.e. patient signals have to be bunched in equilibrium.

Fourth, in a partition equilibrium communicating a higher partition will result in a higher treatment decision. This follows from the fact that higher signals σ^p indicate higher health states θ and the doctor's optimal treatment decision is increasing in θ. Formally speaking, $H(\theta|(s_1, s_2), \sigma^d)$ first order stochastically dominates $H(\theta|(s_1', s_2'), \sigma^d)$

$H(0|0,0) < 1$. A patient will then always desire a treatment that is strictly bounded away from 0. Therefore, interests of patient and doctor are not aligned even if the constraint $\tau \geq 0$ is binding.

[115] To be precise, this follows as the treatment range is bounded by $\underline{\tau}$ and 1. On this closed and bounded treatment range the maximum of the second derivative exists and constitutes the bound away from 0.

whenever $s_1' < s_2' \leq s_1 < s_2$.

Fifth, in a partition equilibrium there exists a minimum partition length $\kappa > 0$. It was shown earlier that the optimal treatment decision of a doctor is uniform continuous in σ^p (in a hypothetical separating equilibrium). Therefore, there exists a $\kappa > 0$ such that optimal treatment decisions differ by less than b for all σ^p and $\sigma^{p\prime}$ with $|\sigma^p - \sigma^{p\prime}| < \kappa$ (in a hypothetical separating equilibrium). Now suppose by way of contradiction that there was a partition (s_0, s_1) with $s_1 - s_0 < \kappa$. By the definition of κ and b, a patient with signal $\sigma^p = s_0$ will (in expectation) strictly prefer the cost sensitive doctor's separating treatment decision for type $\sigma^p = s_1$ to the separating treatment decision for type $\sigma^p = s_0$. By concavity of $u(\cdot)$, he will also prefer a cost sensitive doctor's separating treatment decision for all types $\sigma^p \in (s_0, s_1)$ to his own. By continuity, the same holds for patients with a signal $s_0 - \varepsilon$ for some $\varepsilon > 0$ small enough. Clearly, a cost sensitive doctor receiving the message (s_0, s_1) will assign a treatment between the optimal separating treatment for $\sigma^p = s_0$ and for $\sigma^p = s_1$. Therefore, a patient with signal $s_0 - \varepsilon$ will prefer the message (s_0, s_1) to any message $m \subset [0, s_0]$.

Step five and boundedness of the patient's signal range imply that the number of partitions in any partition equilibrium is bounded.

A one-part-partition equilibrium ("babbling equilibrium") in which all σ^p are pooled exists always. This proves existence of partition equilibria. *Q.E.D.*

Proof of proposition 5.2: Denote the doctor's beliefs over states θ (derived by Bayes' rule) given a signal drawn from $\varepsilon G(\sigma^p, \sigma^d|\theta) + (1 - \varepsilon)1_\theta$ by $k(\theta, \varepsilon|\sigma^d)$. Note that these beliefs are continuous in ε. For $\varepsilon = 0$, the doctor has full information and therefore the welfare maximum is attained with cost incentives. As $c'(\tau) > 0$, decisions under no cost incentives differ from decisions with cost incentives. Consequently, welfare with cost incentives is strictly higher than without cost incentives if $\varepsilon = 0$. As beliefs (and therefore treatment decisions and welfare) are continuous in ε, the first part of the proposition follows.

For the second part, note that $H(\theta|\sigma^p, \sigma^d)$ does not depend on σ^p if $\varepsilon = 0$. Consequently, no information is lost when switching to cost incentives. Taking costs into account makes cost incentives strictly superior as $c'(\tau) > 0$. By continuity of $H(\theta|\sigma^p, \sigma^d)$ in ε, the same conclusion holds for $\varepsilon > 0$ small enough. *Q.E.D.*

Proof of proposition 5.3: Since $c'(\tau) \geq \delta > 0$, there exists an α such that

$$-u'(1) - \alpha c'(0) \leq 0.$$

This implies that the welfare maximizing treatment decision τ is non-positive for any signal/message under the cost function $\alpha c(\tau)$. Without cost incentives $\tau \geq 0$ and $\tau(\sigma^p, \sigma^d) > 0$ with strictly positive probability as

$$\int_0^1 -u'(\theta)\, dH(\theta|\sigma^p, \sigma^d) > 0$$

whenever $H(0|\sigma^p, \sigma^d) < 1$. Consequently, welfare is lower without cost incentives compared to the simple policy $\tau = 0$ (regardless of the signal) under cost function $\alpha c(\tau)$. A cost sensitive doctor will improve on this simple policy by using the information he has, i.e. σ^d. Consequently, cost incentives lead to higher welfare than no cost incentives under the cost function $\alpha c(\tau)$. *Q.E.D.*

Bibliography

Akerlof, G. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics 84*(3), 488–500.

Alonso, R., W. Dessein, and N. Matouschek (2008). When does coordination require centralization? *American Economic Review 98*(1), 145–179.

Araujo, A., L. de Castro, and H. Moreira (2008). Non-monotoniticies and the all-pay auction tie-breaking rule. *Economic Theory 35*(3), 407–440.

Araujo, A., D. Gottlieb, and H. Moreira (2007). A model of mixed signals with applications to countersignalling. *RAND Journal of Economics 38*(4), 1020–1043.

Araujo, A. and H. Moreira (2001). Adverse selection problems without the Spence-Mirrlees condition. mimeo IMPA.

Araujo, A. and H. Moreira (2003). Non-monotone insurance contracts and their empirical counterparts. Discussion Paper #512, Getulio Vargas Foundation.

Araujo, A. and H. Moreira (2010). Adverse selection problems without the Spence-Mirrlees condition. *Journal of Economic Theory 145*(5), 1113–1141.

Araujo, A., H. Moreira, and M. Tsuchida (2007). The trade-off between incentives and endogenous risk. *Brazilian Review of Econometrics 27*(2), 173–198.

Araujo, A., H. Moreira, and M. Tsuchida (2011). Do dividend changes signal future earnings? *Journal of Financial Intermediation 20*(1), 117–134.

Araujo, A., H. Moreira, and S. Vieira (2010). The marginal tariff approach without single-crossing. mimeo IMPA.

Armstrong, M. (1996). Multiproduct nonlinear pricing. *Econometrica 64*(1), 51–75.

Arrow, K. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review 53*(5), 941–973.

Asker, J. and E. Cantillon (2008). Properties of scoring auctions. *RAND Journal of Economics 39*(1), 69–85.

Asker, J. and E. Cantillon (2010). Procurement when price and quality matter. *RAND Journal of Economics 41*(1), 1–34.

Bagnoli, M. and T. Bergstrom (2005). Log-concave probability and its applications. *Economic Theory 26*(2), 445–469.

Bajari, P., H. Hong, and A. Khwaja (2005). A semiparametric analysis of adverse selection and moral hazard in health insurance contracts. Technical report, Preliminary Mimeo, University of Michigan and Duke University.

Baron, D. and R. Myerson (1982). Regulating a monopolist with unknown costs. *Econometrica 50*(4), 911–930.

Baumol, W. J., J. C. Panzar, and R. D. Willig (1982). *Constestable markets and the theory of industry structure*. New York: Harcourt Brace Jovanovich, Inc.

Beard, T., S. Caudill, and D. Gropper (1991). Finite mixture estimation of multiproduct cost functions. *Review of Economics and Statistics*, 654–664.

Billingsley, P. (1986). *Probability and measure*. New York: Wiley, J.

Bolton, P. and M. Dewatripont (2005). *Contract theory*. MIT Press.

Boone, J. and C. Schottmüller (2011a). Health insurance without single crossing: Why healthy people have high coverage. *CEPR Discussion Paper no. 8501*.

Boone, J. and C. Schottmüller (2011b). Procurement with specialized firms. *CEPR Discussion Paper no. 8704*.

Branco, F. (1997). The design of multidimensional auctions. *RAND Journal of Economics 28*(1), 63–81.

Brocas, I. (2011). Countervailing incentives in allocation mechanisms with type-dependent externalities. Discussion Paper.

Calzolari, G. (2004). Incentive regulation of multinational enterprises. *International Economic Review 45*(1), 257–282.

Cardon, J. H. and I. Hendel (2001). Asymmetric information in health insurance: Evidence from the national medical expenditure survey. *RAND Journal of Economics 32*(3), 408–427.

Carrillo, J. (1998). Coordination and externalities. *Journal of Economic Theory 78*(1), 103–129.

Che, Y.-K. (1993). Design competition through multidimensional auctions. *RAND Journal of Economics 24*(4), 668–680.

Chiappori, P.-A., B. Jullien, B. Salanié, and F. Salanié (2006). Asymmetric information in insurance: General testable implications. *RAND Journal of Economics 37*(4), 783–798.

Chone, P. and G. Laroque (2010). Negative marginal tax rates and heterogeneity. *American Economic Review 100*(5).

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics 3*(1), 1–44.

Cohn, J. (2007). *Sick: The untold story of America's health care crisis–and the people who pay the price*. Harper Perennial.

Crawford, V. and J. Sobel (1982). Strategic information transmission. *Econometrica*, 1431–1451.

Cutler, D. M., A. Finkelstein, and K. McGarry (2008). Preference heterogeneity and insurance markets: Explaining a puzzle of insurance. *American Economic Review* (Papers and Proceedings) *98*(2), pp. 157–162.

Dafny, L. (2010). Are health insurance markets competitive? *American Economic Review 100*(4), 1399–1431.

de Barreda, I. (2010). Cheap talk with two-sided private information. London School of Economics; mimeo.

De Meza, D. and D. Webb (2001). Advantageous selection in insurance markets. *RAND Journal of Economics 32*(2), 249–262.

Dupuit, J. (1849). On tolls and transport charges. translated in International Economics Papers (London: Macmillan, 1962). Original version in Annales des Ponts et Chaussées 17 (1849). *International Economic Papers 11*, 7–31.

Einav, L. and A. Finkelstein (2011). Selection in insurance markets: Theory and empirics in pictures. *Journal of Economic Perspectives 25*(1), 115–138.

Emanuel, E. and N. Dubler (1995). Preserving the physician-patient relationship in the era of managed care. *JAMA: Journal of the American Medical Association 273*(4), 323–329.

Fang, H., M. P. Keane, and D. Silverman (2008). Sources of advantageous selection: Evidence from the medigap insurance market. *Journal of Political Economy 116*(2), 303–350.

Figueroa, N. and V. Skreta (2009). The role of optimal threats in auction design. *Journal of Economic Theory 144*(2), 884–897.

Finkelstein, A. and K. McGarry (2006). Multiple dimensions of private information: evidence from the long-term care insurance market. *American Economic Review 96*(4), 938–958.

Frijters, P., J. Haisken-DeNew, and M. Shields (2005). The causal effect of income on health: Evidence from German reunification. *Journal of Health Economics 24*(5), 997–1017.

Fudenberg, D. and J. Tirole (1991). *Game theory*. MIT Press.

Gallagher, T. and W. Levinson (2004). A prescription for protecting the doctor-patient relationship. *American Journal of Managed Care 10*(2; part 1), 61–68.

Gallagher, T., R. St Peter, M. Chesney, and B. Lo (2001). Patients attitudes toward cost control bonuses for managed care physicians. *Health Affairs 20*(2), 186–192.

García, D. (2005). Monotonicity in direct revelation mechanisms. *Economics Letters 88*(1), 21–26.

Glicksberg, I. L. (1952). A further generalization of the Kakutani fixed point theorem with application to Nash equilibrium points. *Proceedings of the American Mathematical Society 38*, 170–174.

Glied, S. (2000). Managed care. Volume 1, chapter 13 of *Handbook of Health Economics*, pp. 707–753. Elsevier.

Goldman, D., G. Joyce, and Y. Zheng (2007). Prescription drug cost sharing: associations with medication and medical utilization and spending and health. *JAMA: Journal of the American Medical Association 298*(1), 61–69.

Gravelle, H. and M. Sutton (2009). Income, relative income, and self-reported health in Britain 1979-2000. *Health Economics 18*(2), 125–145.

Guasch, J. L. and A. Weiss (1981). Self-selection in the labor market. *American Economic Review 71*(3), 275–284.

Guesnerie, R. and J.-J. Laffont (1984). A complete solution to a class of principal-agent problems with an application to the control of a self-managed firm. *Journal of Public Economics 25*(3), 329 – 369.

Harsanyi, J. (1967). Games with incomplete information played by "Bayesian" players, i-iii. part i. the basic model. *Management Science 14*(3), 159–182.

Hellwig, M. (2010). Incentive problems with unidimensional hidden characteristics: A unified approach. *Econometrica 78*(4), 1201–1237.

Hemenway, D. (1990). Propitious selection. *Quarterly Journal of Economics 105*(4), 1063–1069.

Hoffmann, F. and R. Inderst (2011). Pre-sale information. *Journal of Economic Theory 146*(6), 2333–2355.

Jack, W. (2006). Optimal risk adjustment with adverse selection and spatial competition. *Journal of Health Economics 25*(5), 908 – 926.

Jullien, B. (2000). Participation constraints in adverse selection models. *Journal of Economic Theory 93*(1), 1–47.

Jullien, B., B. Salanie, and F. Salanie (2007). Screening risk-averse agents under moral hazard: Single-crossing and the CARA case. *Economic Theory 30*(1), 151–169.

Kao, A., D. Green, A. Zaslavsky, J. Koplan, and P. Cleary (1998). The relationship between method of physician payment and patient trust. *JAMA: Journal of the American Medical Association 280*(19), 1708–1714.

Kerr, E., R. Hays, B. Mittman, A. Siu, B. Leake, and R. Brook (1997). Primary care physicians' satisfaction with quality of care in california capitated medical groups. *JAMA: Journal of the American Medical Association 278*(4), 308–312.

Laffont, J.-J. and J. Tirole (1987). Auctioning incentive contracts. *Journal of Political Economy 95*(5), 921–937.

Laffont, J.-J. and J. Tirole (1993). *A theory of incentives in procurement and regulation*. MIT Press.

Lewis, T. and D. Sappington (1989). Countervailing incentives in agency problems. *Journal of Economic Theory 49*(2), 294–313.

Ma, C. and T. McGuire (1997). Optimal health insurance and provider payment. *American Economic Review 87*(4), 685–704.

Martimort, D. and L. Stole (2009). Market participation in delegated and intrinsic common-agency games. *RAND Journal of Economics 40*(1), 78–102.

Maskin, E. and J. Riley (1984). Optimal auctions with risk averse buyers. *Econometrica 52*(6), 1473–1518.

Matthews, S. and J. Moore (1987). Monopoly provision of quality and warranties: An exploration in the theory of multidimensional screening. *Econometrica 55*(2), 441–467.

McAfee, R. and J. McMillan (1988). Multidimensional incentive compatibility and mechanism design. *Journal of Economic Theory 46*(2), 335–354.

Mechanic, D. and M. Schlesinger (1996). The impact of managed care on patients' trust in medical care and their physicians. *JAMA: Journal of the American Medical Association 275*(21), 1693–1697.

Miravete, E. and L. Röller (2004). Estimating price-cost markups under nonlinear pricing competition. *Journal of the European Economic Association 2*(2-3), 526–535.

Mirrlees, J. (1971). An exploration in the theory of optimum income taxation. *Review of Economic Studies 38*(2), 175–208.

Munkin, M. and P. Trivedi (2010). Disentangling incentives effects of insurance coverage from adverse selection in the case of drug expenditure: A finite mixture approach. *Health Economics 19*(9), 1093–1108.

Mussa, M. and S. Rosen (1978). Monopoly and product quality. *Journal of Economic Theory 18*(2), 301–317.

Myerson, R. (1979). Incentive compatibility and the bargaining problem. *Econometrica 47*, 61–73.

Myerson, R. (1981). Optimal auction design. *Mathematics of Operations Research 6*, 58–73.

Netzer, N. and F. Scheuer (2010). Competitive screening in insurance markets with endogenous wealth heterogeneity. *Economic Theory 44*(2), 187–211.

Olivella, P. and M. Vera-Hernández (2007). Competition among differentiated health plans under adverse selection. *Journal of Health Economics 26*(2), 233 – 250.

Pauly, M. (1968). The economics of moral hazard: Comment. *American Economic Review 58*(3), 531–537.

Piette, J., M. Heisler, and T. Wagner (2004a). Cost-related medication underuse among chronically ill adults: The treatments people forgo, how often, and who is at risk. *American Journal of Public Health 94*(10), 1782–1787.

Piette, J., M. Heisler, and T. Wagner (2004b). Problems paying out-of-pocket medication costs among older adults with diabetes. *Diabetes Care 27*(2), 384–391.

Rochet, J. (2009). Monopoly regulation without the Spence-Mirrlees assumption. *Journal of Mathematical Economics 45*(9-10), 693–700.

Rochet, J. and P. Choné (1998). Ironing, sweeping, and multidimensional screening. *Econometrica 66*(4), 783–826.

Rochet, J. and L. Stole (2003). The economics of multidimensional screening. In *Advances in economics and econometrics: Theory and applications, Eighth World Congress*, Volume 1, pp. 150–197. Cambridge University Press.

Rodwin, M. (1995). Conflicts in managed care. *New England Journal of Medicine 332*(9), 604–607.

Rothschild, M. and J. Stiglitz (1976). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. *Quarterly Journal of Economics 90*(4), 629–649.

Schoen, C., S. R. Collins, J. L. Kriss, and M. M. Doty (2008). How many are underinsured? trends among U.S. adults, 2003 and 2007. *Health Affairs (Project Hope) 27*(4), 298–309.

Schoen, C., R. Osborn, D. Squires, M. M. Doty, R. Pierson, and S. Applebaum (2010). How health insurance design affects access to care and costs, by income, in eleven countries. *Health Affairs 29*(12), 1–12.

Schottmüller, C. (2011a). Adverse selection without single crossing: The monotone solution. *Center Discussion Paper, no. 2011–123*.

Schottmüller, C. (2011b). Cost incentives for doctors: A double-edged sword. *Tilec Discussion Paper, no. 2011-041*.

Smart, M. (2000). Competitive insurance markets with two unobservables. *International Economic Review 41*(1), 153–170.

Stewart, M. (1995). Effective physician-patient communication and health outcomes: A review. *CMAJ: Canadian Medical Association Journal 152*(9), 1423–1433.

Stiglitz, J. (1977). Monopoly, non-linear pricing and imperfect information: The insurance market. *Review of Economic Studies 44*(3), 407–430.

Stiglitz, J. E. (1982). Self-selection and Pareto efficient taxation. *Journal of Public Economics 17*(2), 213 – 240.

Tirole, J. (1988). *The theory of industrial organization.* MIT Press.

Wambach, A. (2000). Introducing heterogeneity in the Rothschild-Stiglitz model. *Journal of Risk and Insurance 67*(4), 579–591.